Human-like Learning

Peng Cui

Tsinghua University

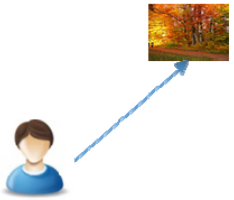
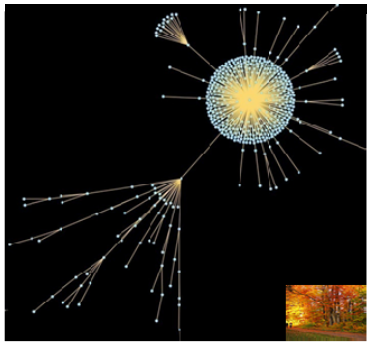
About Me

- Bio
  - Associate Professor in DCS
  - 2009 PhD
- Research interests:
  - Human-like visual learning
  - Causal analysis
  - Network embedding
  - Social dynamics modeling
- More on

  http://pengcui.thumedialab.com

Research: Social Behavioral Modeling

User behavior analysis

Microscopic	Mesoscopic	Macroscopic
		
Individual	**Group**	**Global**
Individual behavior analysis and prediction	Group behaviors/responses analysis and prediction	Propagation analysis and prediction
KDD13, KDD14, KDD15, AAAI15	*KDD16, SIGIR11, AAAI11, AAAI15*	*KDD13, ICDM15, KDD16*
KDD14 Best Paper Finalist	KDD16 Best Paper Finalist	ICDM15 Best Student Paper

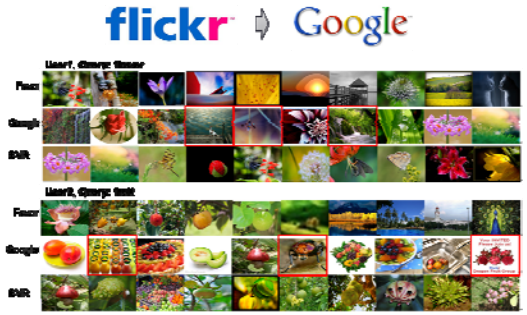
Research: Social-sensed Multimedia Computing

Microscopic

Personalized

Image Search

Social-Sensed Image Search

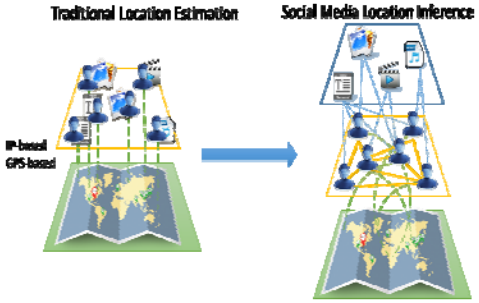


Mesoscopic

Group

LBS

Social-Sensed Location Inference



Macroscopic

Global

Multimedia Summary

Social-Sensed Summarization

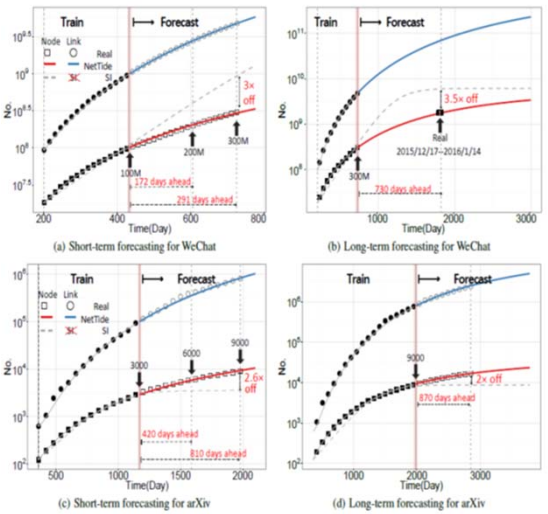
MMM'13 Best Paper Award

ICME'14 Best Paper Award

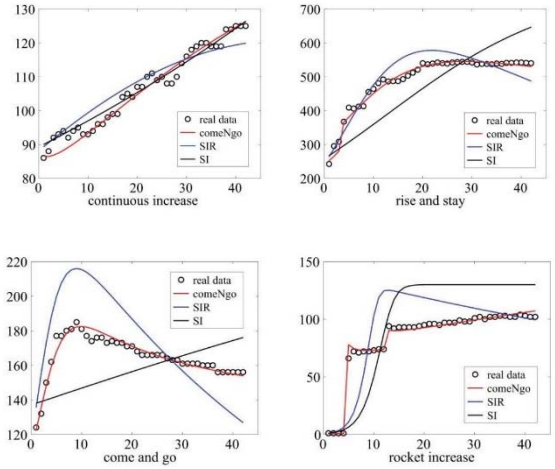
SIGMM'12 Grand Challenge Award

Research：Social Dynamics Modeling

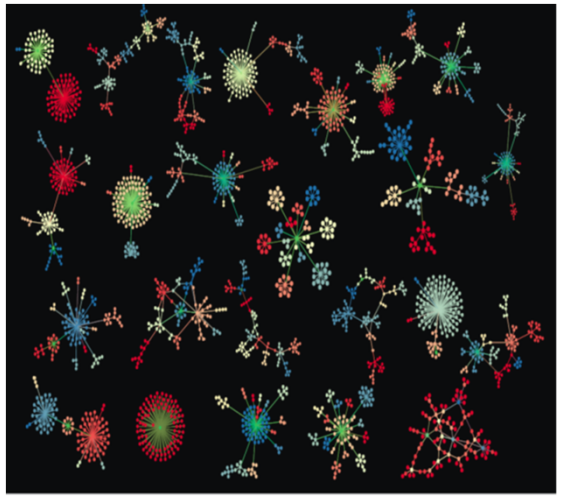
Social network growth



Social group evolution



Information diffusion

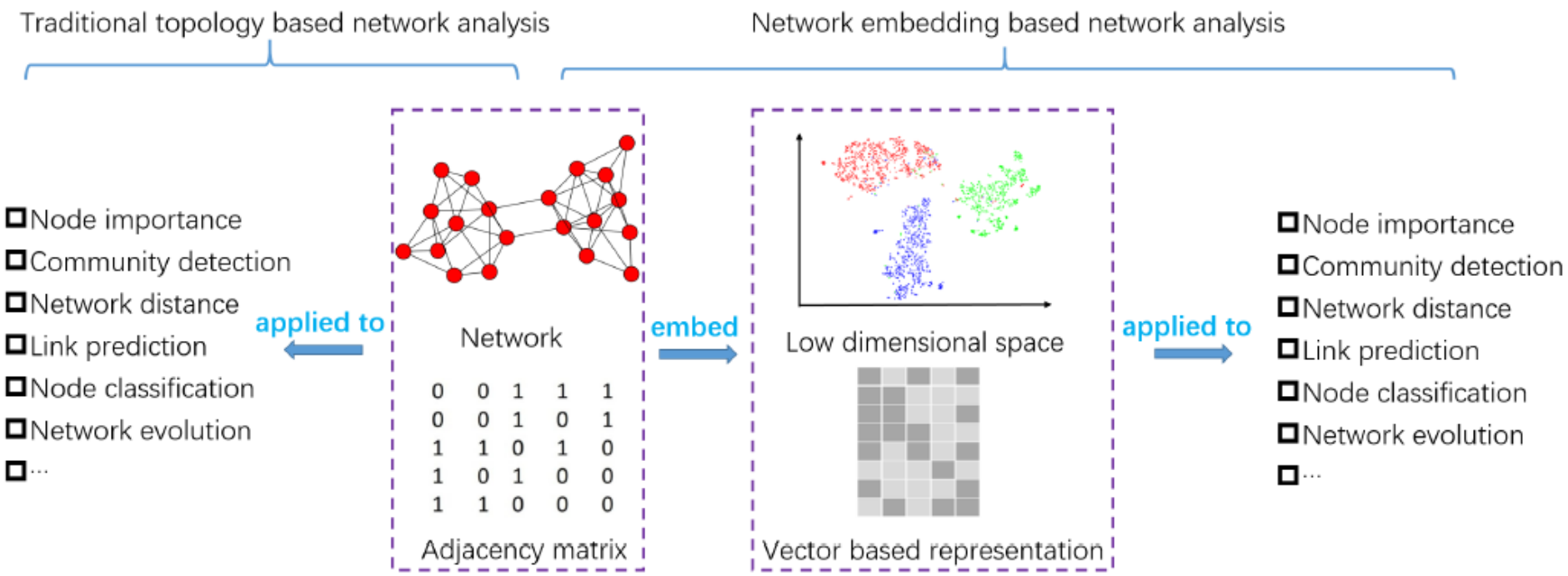


KDD16 Best Paper Finalist

ICDM15 Best Student Paper

Published papers in *KDD16 (2), KDD17 (2), TKDE, PLOS ONE etc.*

Research：Network Embedding

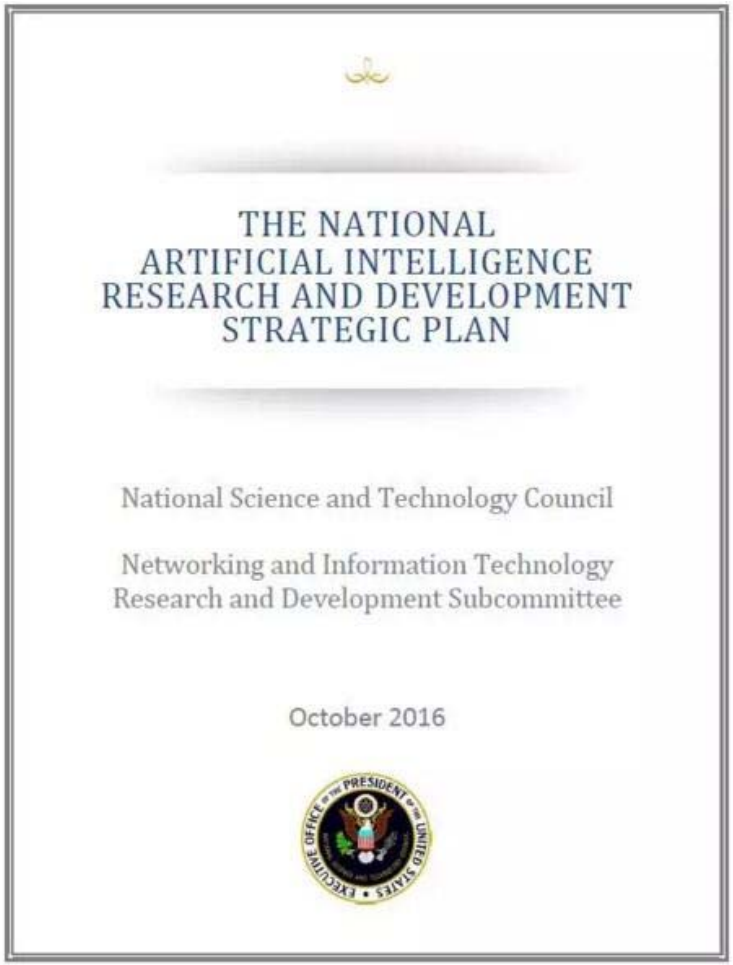


Published papers in *KDD13, KDD 15, KDD16 (2), KDD17 (2), KDD18(6) etc.*
Gave tutorials in *KDD17, AAAI 18, ICDM16, etc.*

Outline

- **Motivation of human-like learning**
- Learning to learn (association)
- Causal inference
- Stable learning
- The NICO dataset

National Strategies for AI

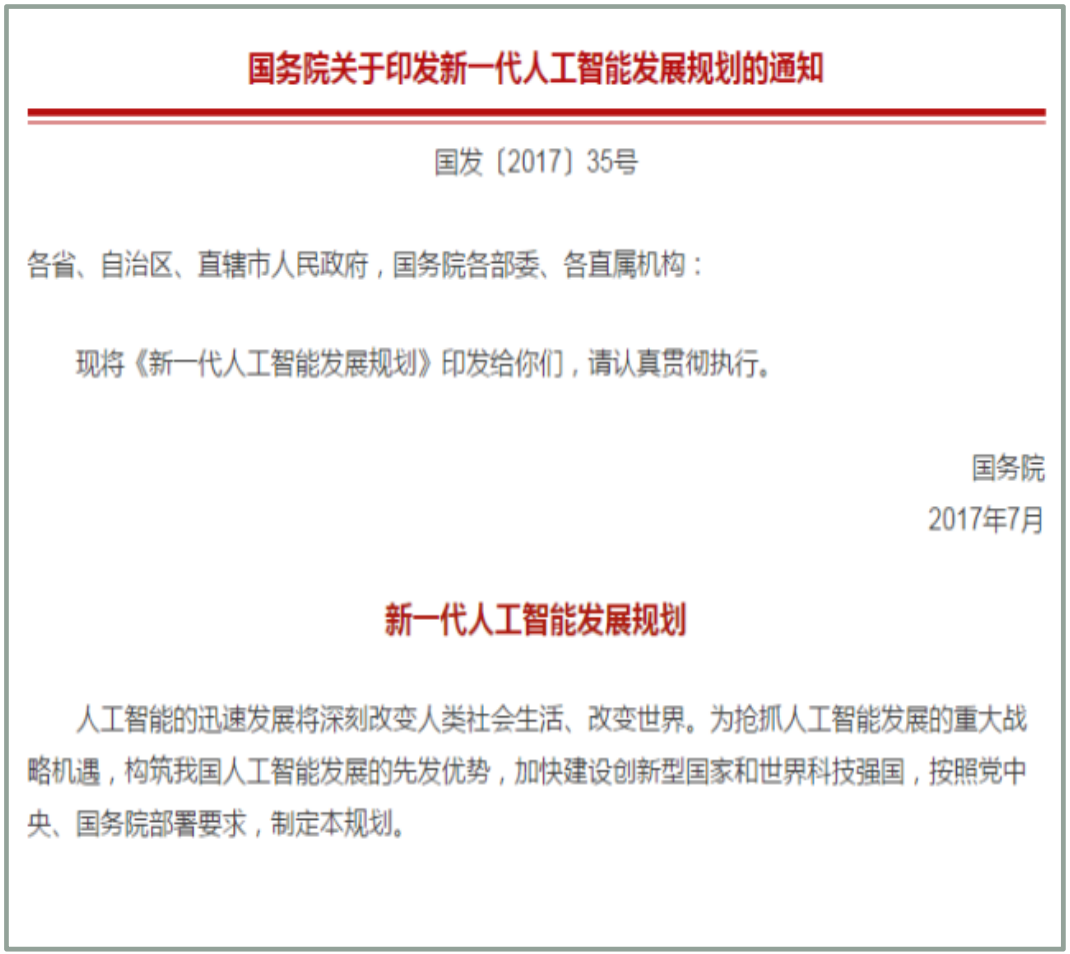


THE NATIONAL
ARTIFICIAL INTELLIGENCE
RESEARCH AND DEVELOPMENT
STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology
Research and Development Subcommittee

October 2016



国务院关于印发新一代人工智能发展规划的通知

国发〔2017〕35号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

现将《新一代人工智能发展规划》印发给你们，请认真贯彻执行。

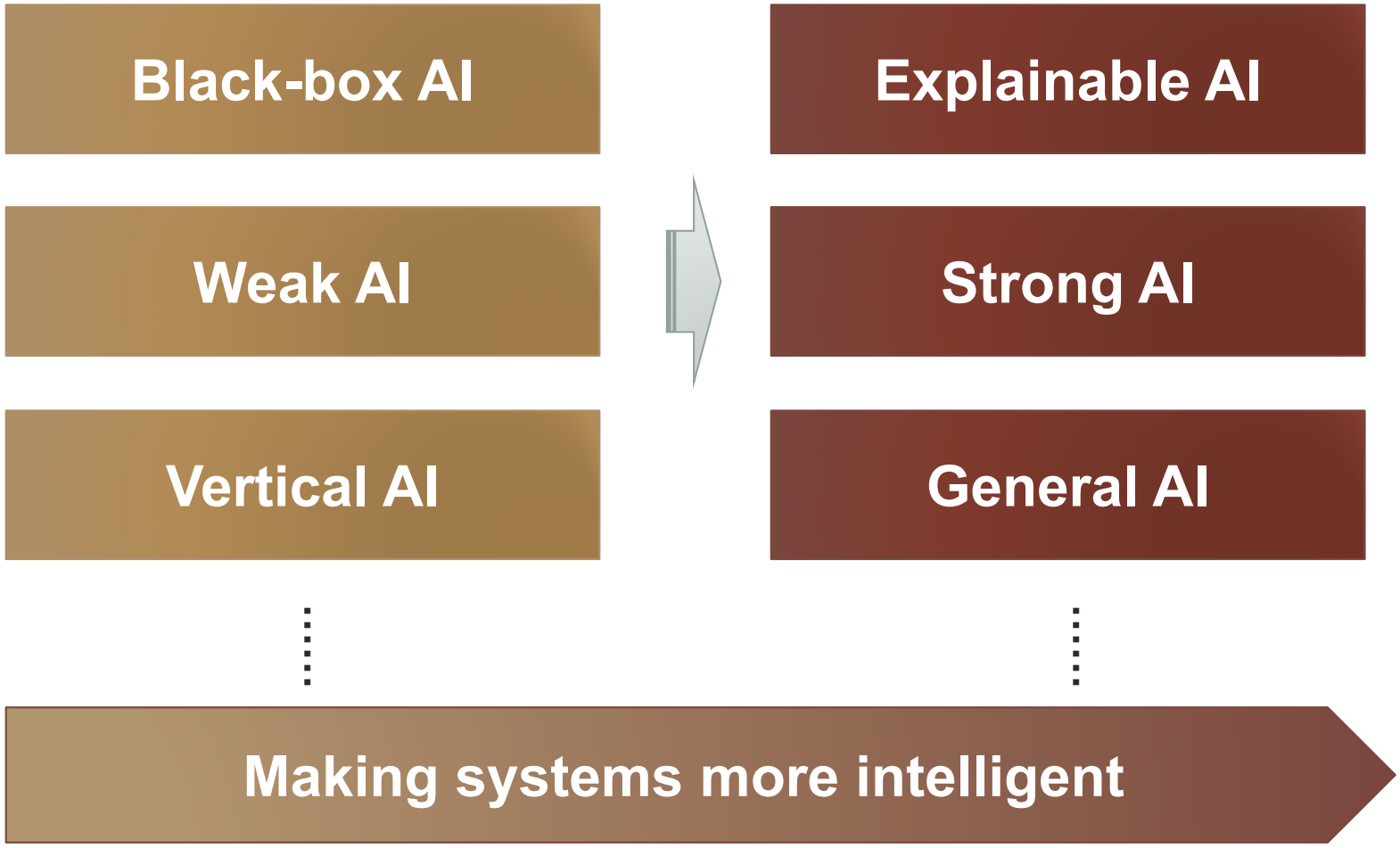
国务院

2017年7月

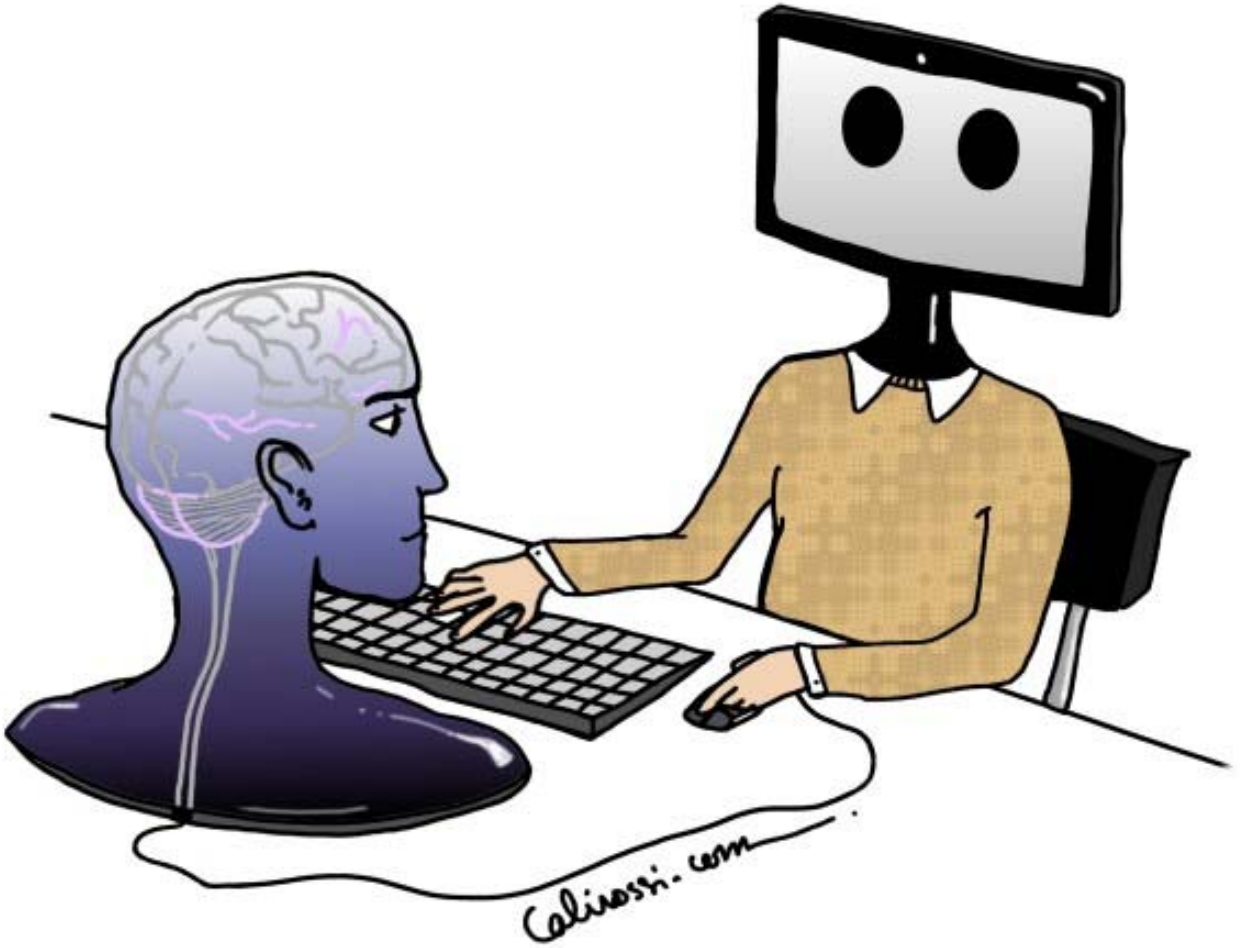
新一代人工智能发展规划

人工智能的迅速发展将深刻改变人类社会生活、改变世界。为抢抓人工智能发展的重大战略机遇，构筑我国人工智能发展的先发优势，加快建设创新型国家和世界科技强国，按照党中央、国务院部署要求，制定本规划。

Towards AI 2.0



Reference Model: Human



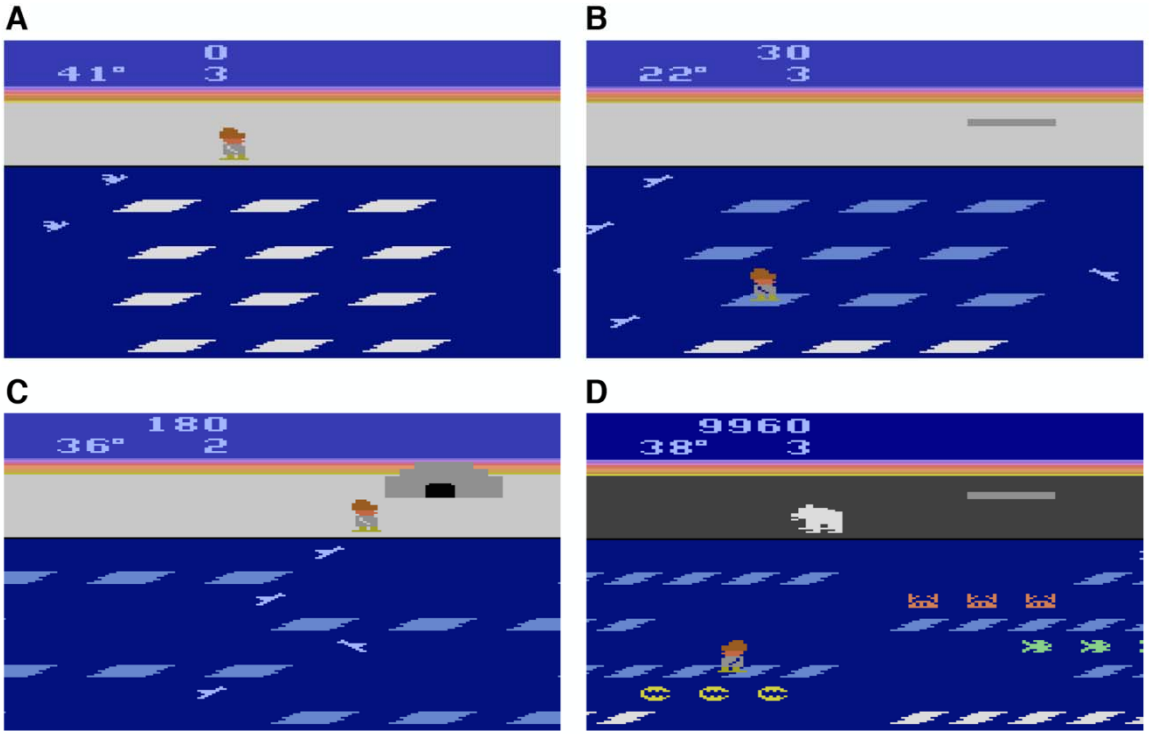
Machine Learning v.s. Human Learning

- Learning Simple Visual Concepts
  - People learn from fewer examples
  - People learn richer representations
  - People can learn to recognize a new character from a single example
  - People learn a concept – a model of the class that allows their acquired knowledge to be flexible applied in new ways.

Lake, Brenden M., et al. “Building machines that learn and think like people.” *Behavioral and Brain Sciences* (2016): 1-101.

Machine Learning v.s. Human Learning

- The Frostbite Challenge



Lake, Brenden M., et al. “Building machines that learn and think like people.” *Behavioral and Brain Sciences* (2016): 1-101.

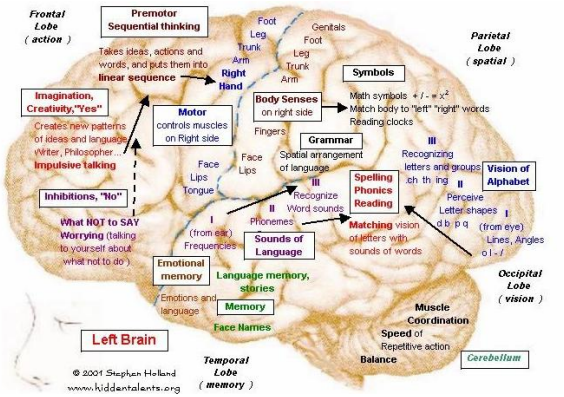
Machine Learning v.s. Human Learning

- The Frostbite Challenge
  - Optimal Solution up till now: Deep Q Network
  - Shortcomings compared with humans
    - People use less time to practice to reach nearly the same average score: human for 2 hours and DQN for 924 hours.
    - Human could grasp the basics of the game just after a few minutes of playing.
    - If humans are able to watch an expert playing for a few minutes, they can learn even faster.
    - Humans are more flexible, i.e. after they learn how to play, they could finish arbitrary new tasks and goals. (e.g. get closest to score 300 etc.)

Lake, Brenden M., et al. “Building machines that learn and think like people.” *Behavioral and Brain Sciences* (2016): 1-101.

Technical Paths of Learning from Human

Brain-like Learning



Bottom-Up

New AI algorithms

Depends on how much we can understand human brain.
The computer architecture v.s. brain architecture?

Technical Paths of Learning from Human

Human-like Learning

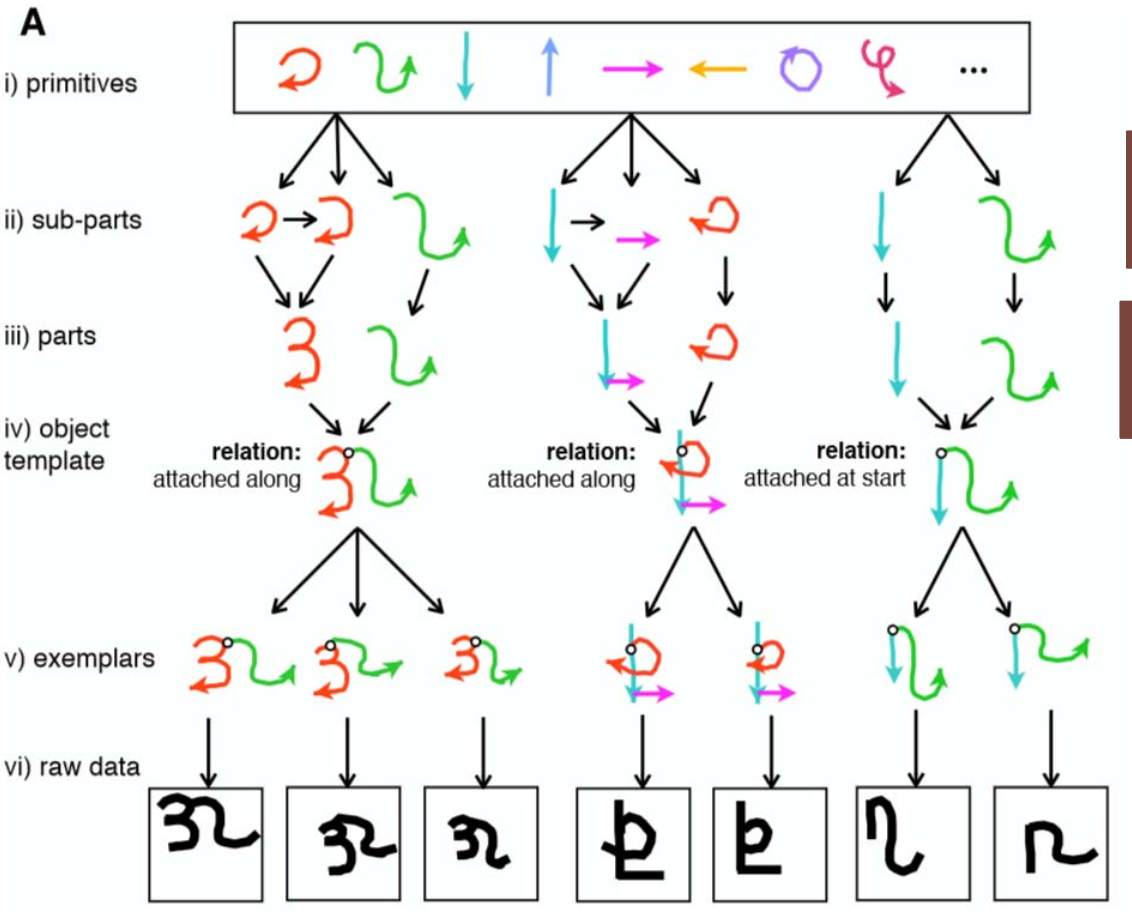
Inference
Association
Imagination

Top-down

New AI algorithms

Another path for inventing new learning mechanisms.

A Successful Example



Inference: compositionality and causality

Association: Learning to learn

Far less samples are required for training than deep models.

Simple showcase

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science, 350 (6266), 1332-1338.

Outline

- Motivation of human-like learning
- **Learning to learn (association)**
- Causal inference
- Stable learning
- The NICO dataset

The problems of today's visual learning

Inept learning way

dog

wolf

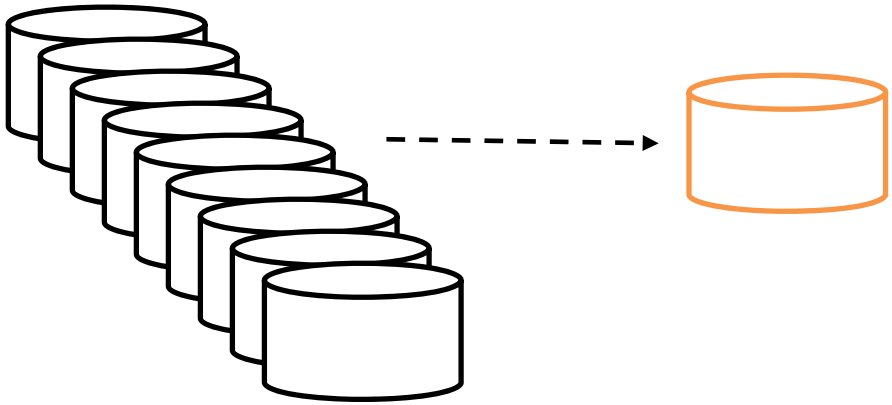
What is the correct way of learning to recognize wolf, given that you have been able to recognize dog?

The problems of today's visual learning

Inept learning way

-How do we **Human** learn a new concept?

-We learn by **association**.



Base Classes Target Class

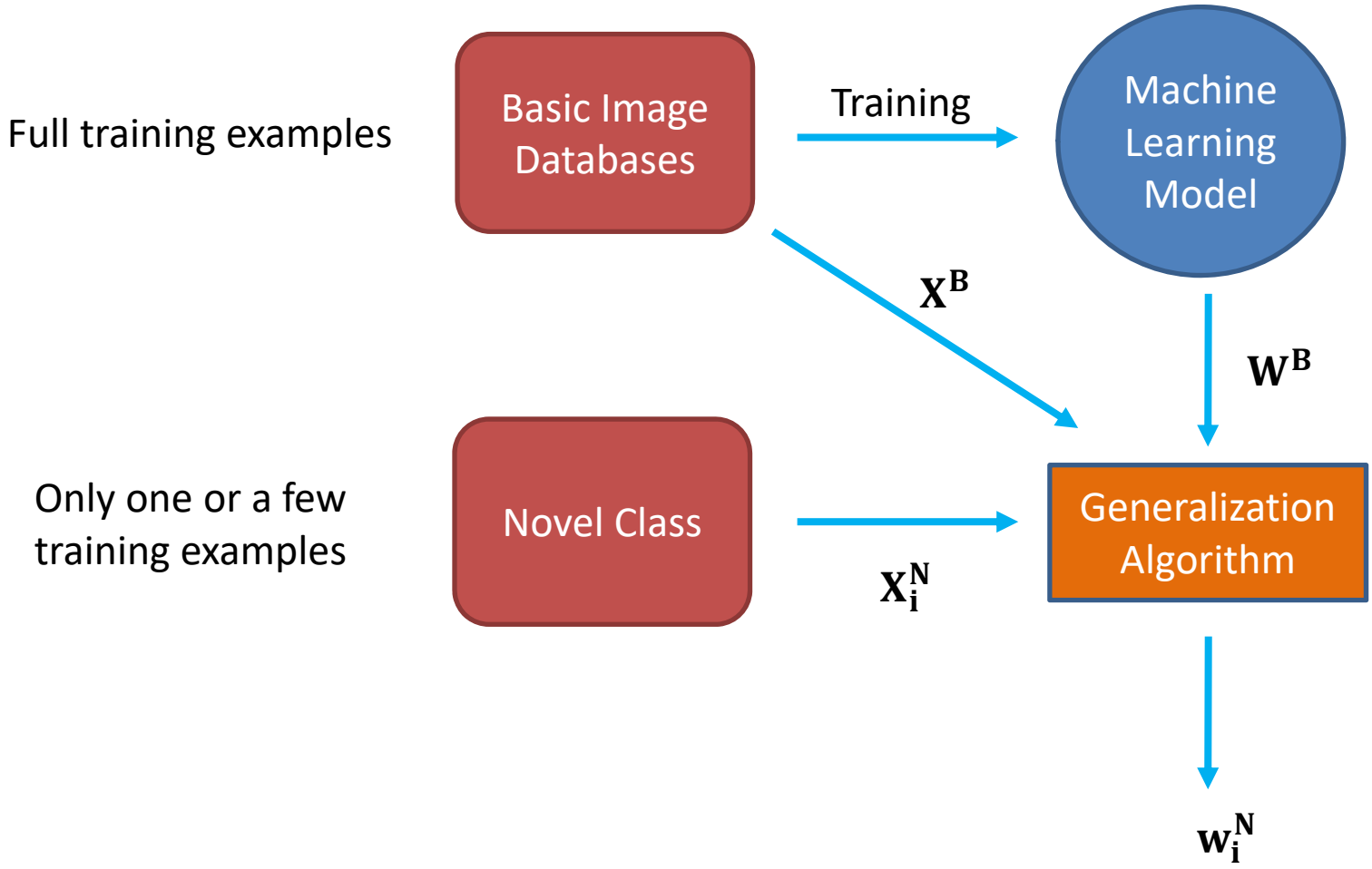
The more we have learned, the faster we should be able to learn new things.

Learning to learn new concepts

Problem Definition

PROBLEM 1 (LEARNING TO LEARN IMAGE CLASSIFIERS). ***Given*** *the image features of base classes* $\mathbf{X}^B$*, the well-trained base classifier parameters* $\mathbf{W}^B$*, and the image features of a novel class* i $\mathbf{X}_i^N$ *with only a few positive samples,* ***learn*** *the classification parameters* $\mathbf{w}_i^N$ *for the novel class, so that the learned classifier* $f(\cdot; \mathbf{w}_i^N | \mathbf{X}^B, \mathbf{W}^B, \mathbf{X}_i^N)$ *can precisely predict labels for the* i^{th} *novel class.*

Problem Definition



Algorithm –VANER (Intuition)

How do human learn a concept without seeing many photos?

Dog
(Base)

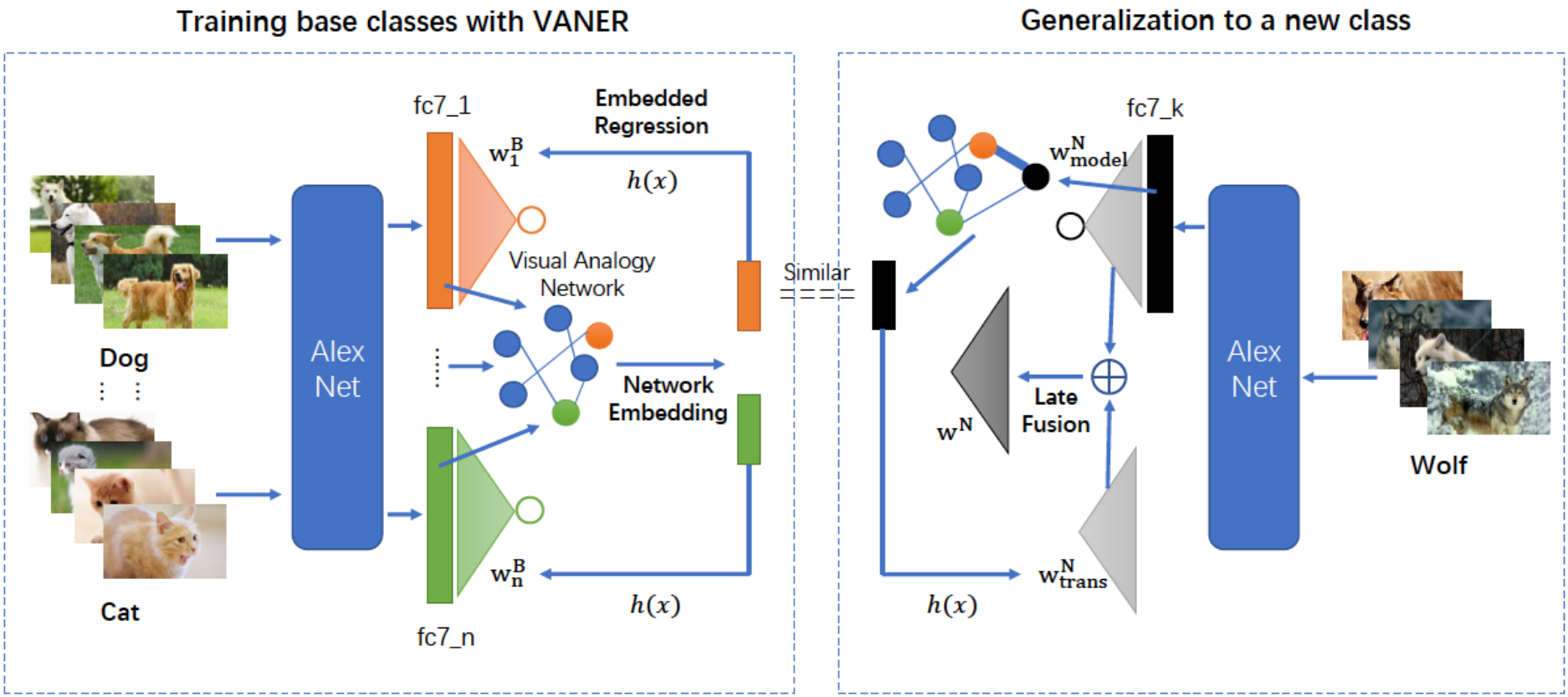
Wolf
(One shot)

Visual Analogy Network

Other concepts…

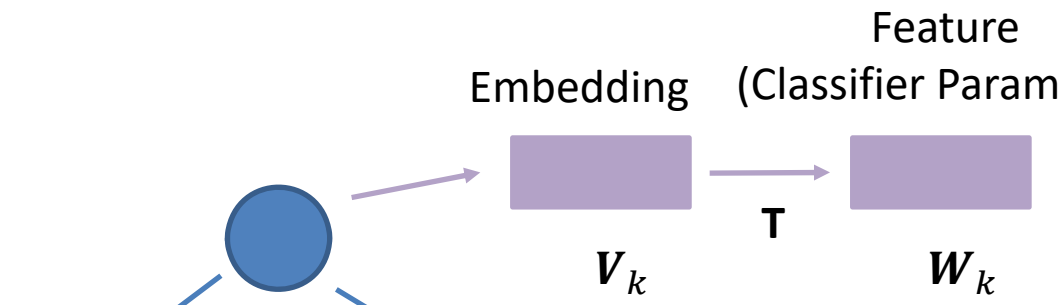
Algorithm – VANER

VANER: Visual Analogy Network Embedded Regression



Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Algorithm – VANER (Details)



A: Similarity Matrix

Training for Base:

Loss Function:

$$\mathscr{L}(\mathbf{V},\mathbf{T}) = \|\mathbf{VT}-\mathbf{W}\|_F^2 + \lambda\|\mathbf{A}-\mathbf{VV}^\top\|_F^2.$$

Solution (Alternative Coordinate Descent):

$$\begin{cases} \dfrac{\partial \mathscr{L}(\mathbf{V},\mathbf{T})}{\partial \mathbf{V}} = 2(\mathbf{VT}-\mathbf{W})\mathbf{T}^\top + \lambda(-4\mathbf{AV}+4\mathbf{VV}^\top\mathbf{V}) \\ \dfrac{\partial \mathscr{L}(\mathbf{V},\mathbf{T})}{\partial \mathbf{T}} = 2\mathbf{V}^\top(\mathbf{VT}-\mathbf{W}). \end{cases}$$

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Algorithm – VANER (Details)

$$\mathscr{L}(\mathbf{V},\mathbf{T}) = \|\mathbf{V}\mathbf{T} - \mathbf{W}\|_F^2 + \lambda\|\mathbf{A} - \mathbf{V}\mathbf{V}^\top\|_F^2.$$

Keeping the precision of the predicted parameter

Keeping the structure of the visual analogy network

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Algorithm – VANER (Late Fusion)

Initializing (w_{trans} as initialization):

$$\mathscr{L}(\mathbf{w}^N) = \left\{ \sum_{\mathbf{x} \in \mathrm{X}_T} L(f(\mathbf{x}, \mathbf{w}^N), y) \right\} + \lambda \cdot R(\mathbf{w}^N), \qquad (8)$$

Tuning:

$$\mathscr{L}(\mathbf{w}^N) = \left\{ \sum_{\mathbf{x} \in \mathrm{X}_T} L(f(\mathbf{x}, \mathbf{w}^N), y) \right\} + \lambda \cdot \left\| \mathbf{w}^N - \mathbf{w}^N_{trans} \right\|_F^2. \qquad (9)$$

Voting (Best):

$$\mathbf{w}^N = \mathbf{w}^N_{trans} + \lambda \cdot \mathbf{w}^N_{model}. \qquad (10)$$

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Experiment Settings

- Dataset: ILSVRC 2015
- 800 Base Classes in ImageNet for training VANER, the base deep network we use is AlexNet
- 200 Novel Classes, each used for binary classification with whole base classes
- For each k-shot problem, we do 10 repeated tests with randomly split in novel class and take the average result.
- Evaluation Metric: AUC / F1 score

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Experiment Baseline

- Logistic Regression (LR)
- Weighted Logistic Regression (Weighted-LR)
- Model Regression Network (MRN)
- VANER
- VANER (-Mapping)
- VANER (-Embedding)

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Experimental Results (1) – Late Fusion

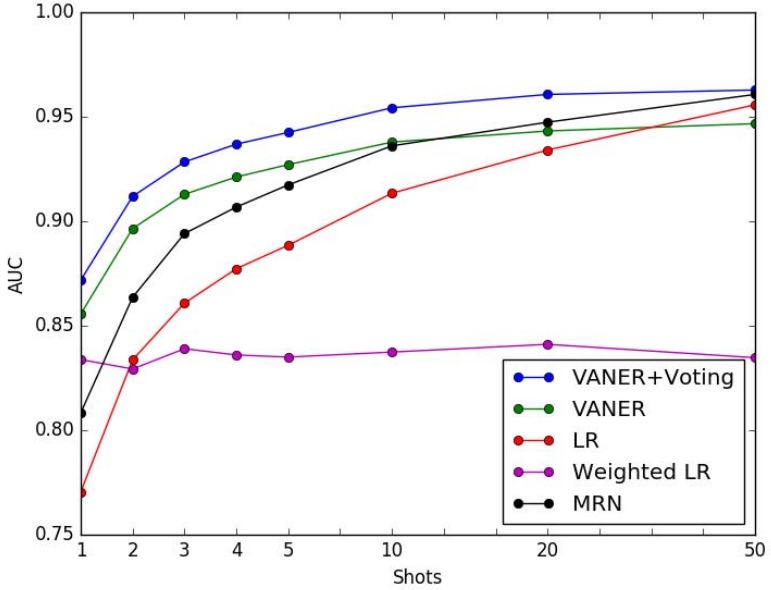
Table 2: Performance of different late fusion mechanism for k-shot problem

Algorithm	1-shot		5-shot		10-shot		20-shot	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
VANER	0.8556	0.5292	0.9271	0.6491	0.9379	0.6721	0.9432	0.6850
VANER + Initializing	0.7662	0.3941	0.9030	0.6185	0.9338	0.6887	0.9461	0.7237
VANER + Tuning	0.7923	0.4244	0.9098	0.6307	0.9365	0.7012	0.9466	0.7268
VANER + Voting	**0.8718**	**0.5671**	**0.9425**	**0.7039**	**0.9543**	**0.7343**	**0.9607**	**0.7510**

The Voting method is proved to be a better method!

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Experimental Results – Algorithm Performance



Compared with logistic regression, we can save 4/5 samples to get similar performance.

Table 1: Performance of different algorithms for k-shot problem

Algorithm	Model Transfer	1-shot		5-shot		10-shot		20-shot	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1
$VANER + Voting^*$	Y	**0.8718**	**0.5671**	**0.9425**	**0.7039**	**0.9543**	**0.7343**	**0.9607**	**0.7510**
$VANER^*$	Y	0.8556	0.5292	0.9271	0.6491	0.9379	0.6721	0.9432	0.6850
$VANER(-Mapping)$	Y	0.8261	0.4551	0.8526	0.4807	0.8726	0.5179	0.8897	0.5394
$VANER(-Embedding)$	Y	0.7922	0.4335	0.9032	0.6015	0.9183	0.6347	0.9393	0.6788
LR	N	0.7705	0.3994	0.8885	0.5882	0.9134	0.6421	0.9341	0.6877
$Weighted - LR$	Y	0.8338	0.4680	0.8350	0.4691	0.8374	0.4711	0.8411	0.4726
MRN	Y	0.8083	0.4511	0.9175	0.6653	0.9361	0.7133	0.9474	0.7388

Experimental Results – Insightful Analysis

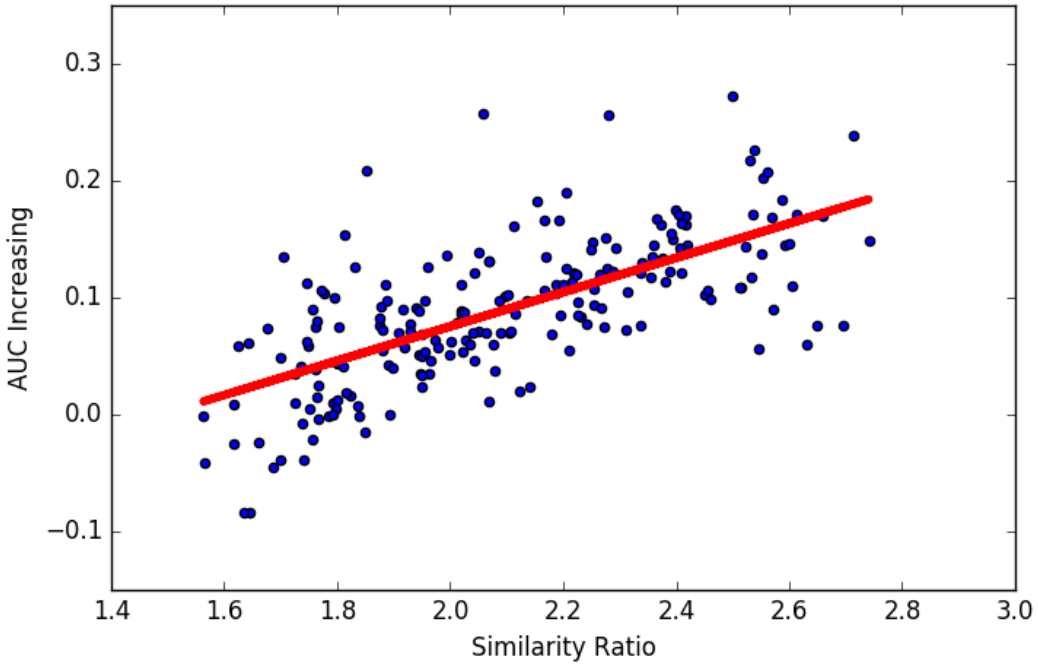
Category	LR (No Transfer)	VANER (Transfer)
Jeep	0.8034	**0.9469**
Zebra	0.8472	**0.9393**
Hen	0.7763	**0.8398**
Lemon	0.6854	**0.9583**
Bubble	**0.7455**	0.7041
Pineapple	0.7364	**0.8623**
Lion	0.8305	**0.9372**
Screen	0.7801	**0.9056**
Drum	0.6510	**0.6995**
Restaurant	0.7806	**0.8787**

Compared with no-transfer algorithm, our VANER is obviously better. However, there are some failure cases like Bubble.

What is the driving factor that controls the success of generalization?

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Experimental Results – Insightful Analysis



$$Def: \quad Similarity\ Ratio = \frac{Average\ Top-k\ Base\ similarity}{Average\ Total\ Base\ similarity}$$

$$AUC\ Increasing = AUC\ for\ VANER - AUC\ for\ LR$$

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, CVPR 2019.

Experimental Results – Embedding Similarity

The embedding layer is explainable:

Novel Class	 Jeep	 Zebra	 Lemon	 Lion	 Screen	 Restaurant
Top-3 Similar Base Classes	 Pickup  Beach wagon  Tow_truck	 Echidna  Leopard  Cheetah	 Orange  Acorn  Granny_Smith	 Cougar  Dingo  Lynx	 Monitor  Laptop  Television	 Shoe_shop  Marimba  Bakery

Figure 3: Top-3 most similar base classes to novel class on embedding layer in 5-shot setting.

Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Informative Visual Analogy, https://arxiv.org/abs/1710.0617

Sectional Summary

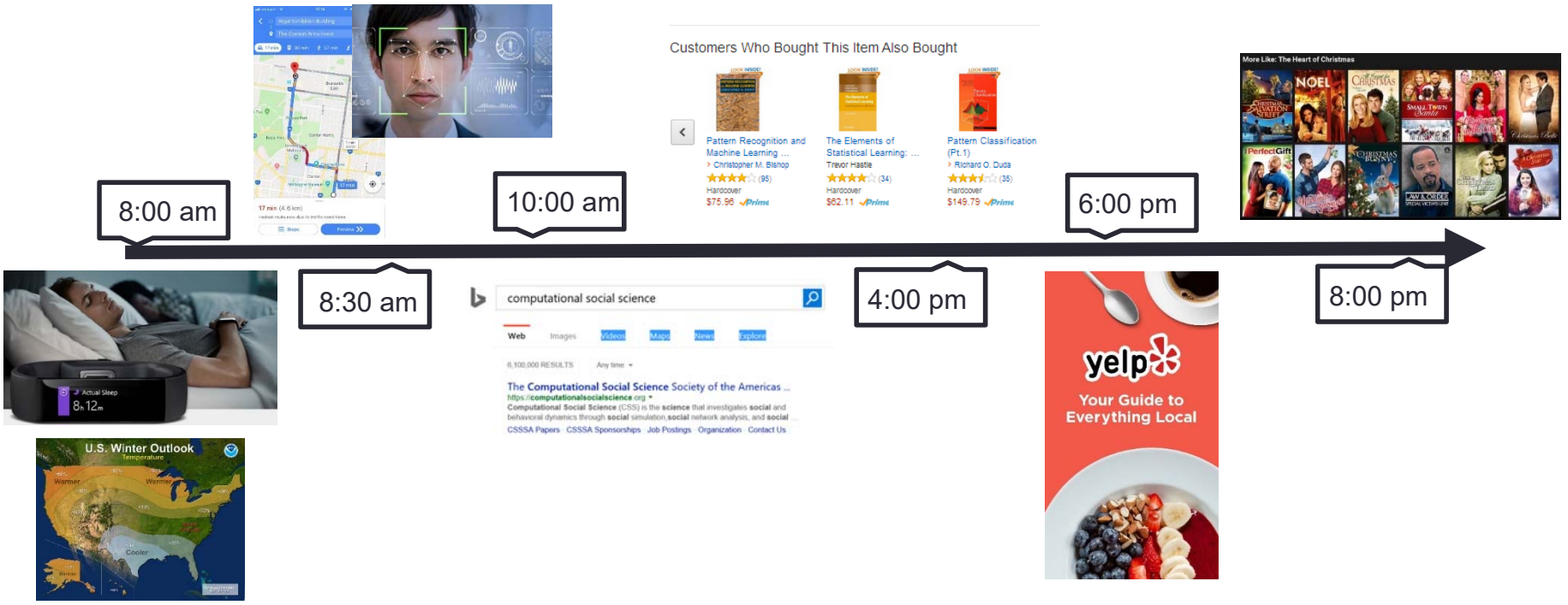
- Small sample learning is a promising direction
- Learning to learn can significantly improve learning efficiency
- From association to imagination?

Outline

- Motivation of human-like learning
- Learning to learn (association)
- **Causal inference**
- Stable learning
- The NICO dataset

ML techniques are impacting our life

- A day in our life with ML techniques



Now we are stepping into risk-sensitive areas

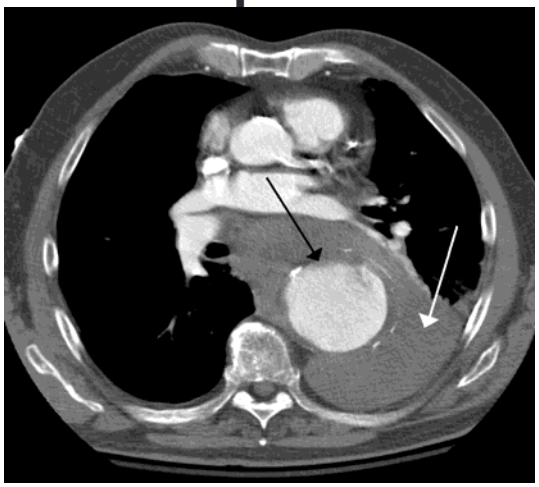
Human

Shifting from ***Performance Driven*** to ***Risk Sensitive***

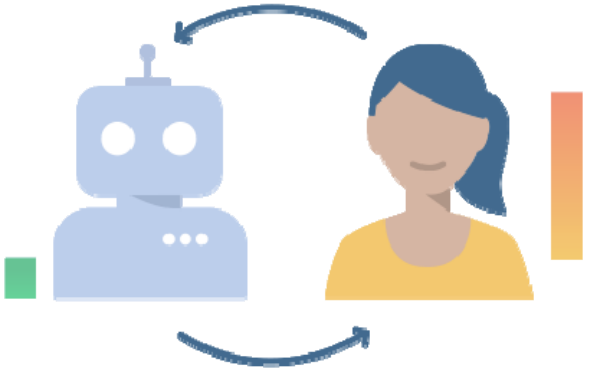
Problems of today's ML - ***Explainability***

Most machine learning models are black-box models

Unexplainable



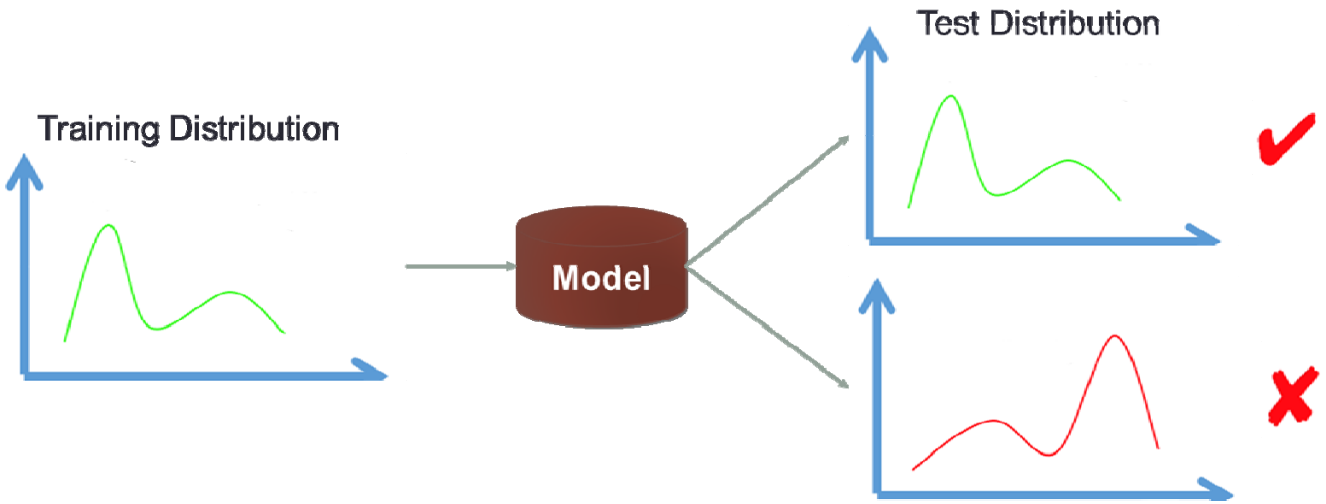
Human in the loop



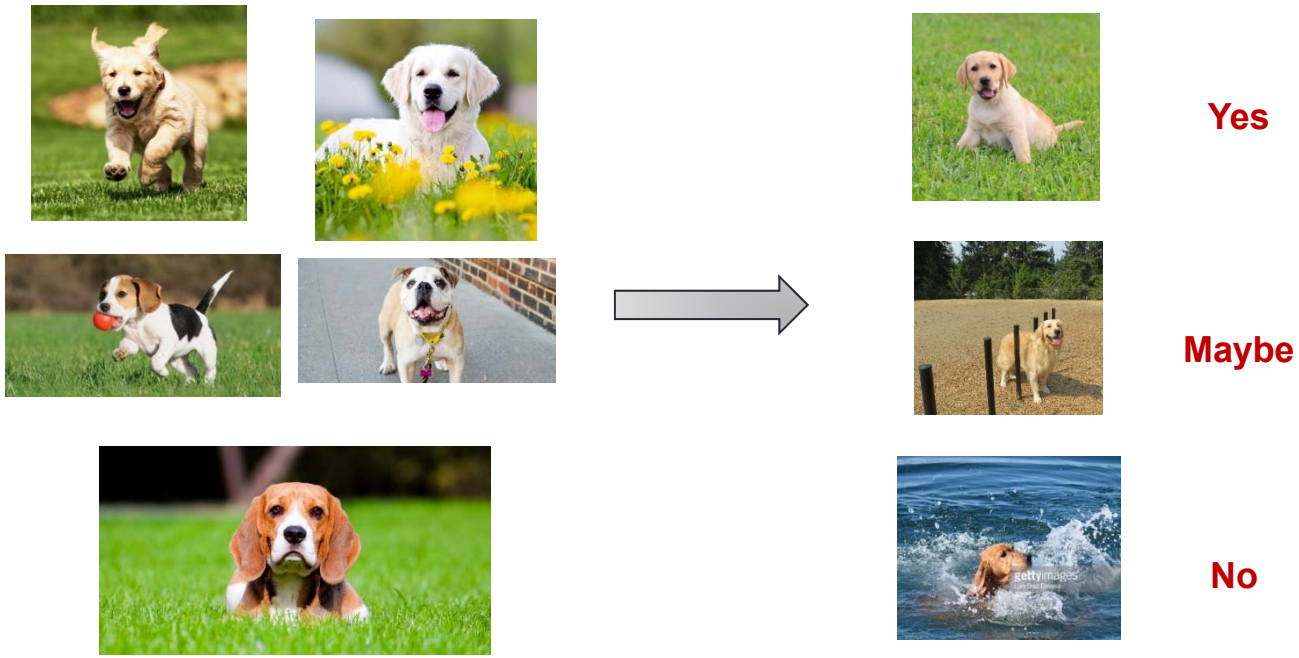
Health Military Finance Industry

Problems of today's ML - ***Stability***

Most ML methods are developed under I.I.D hypothesis

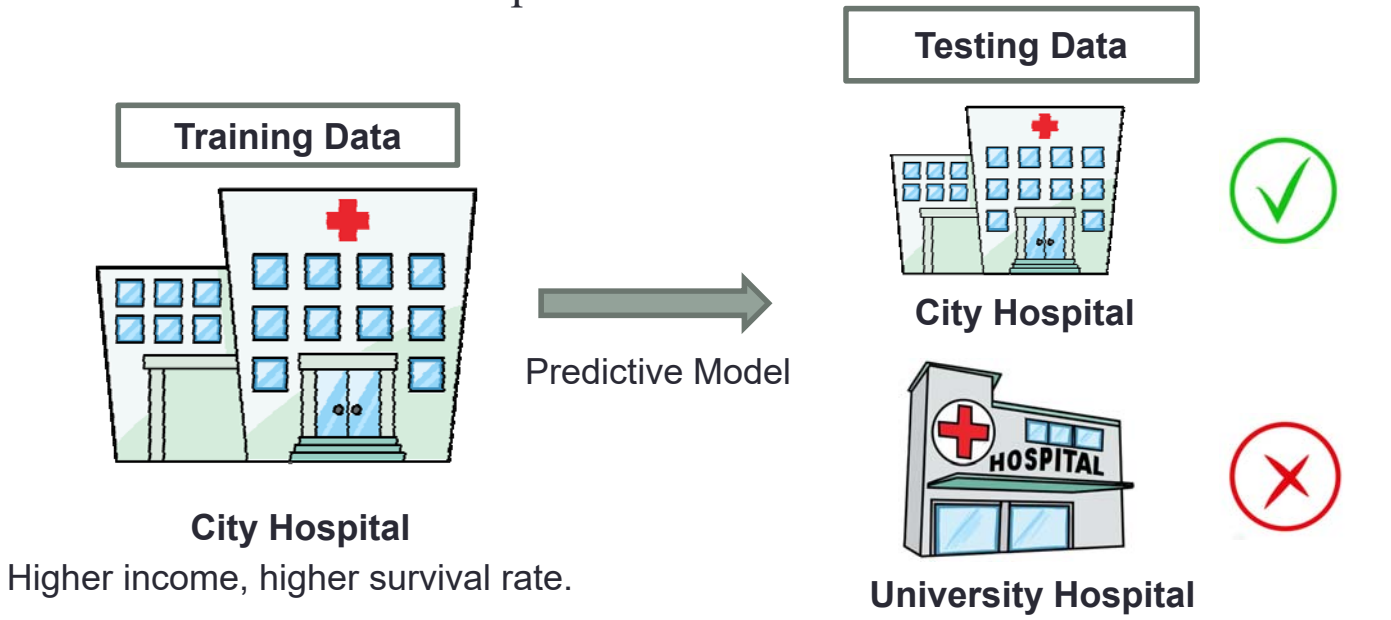


Problems of today's ML - ***Stability***



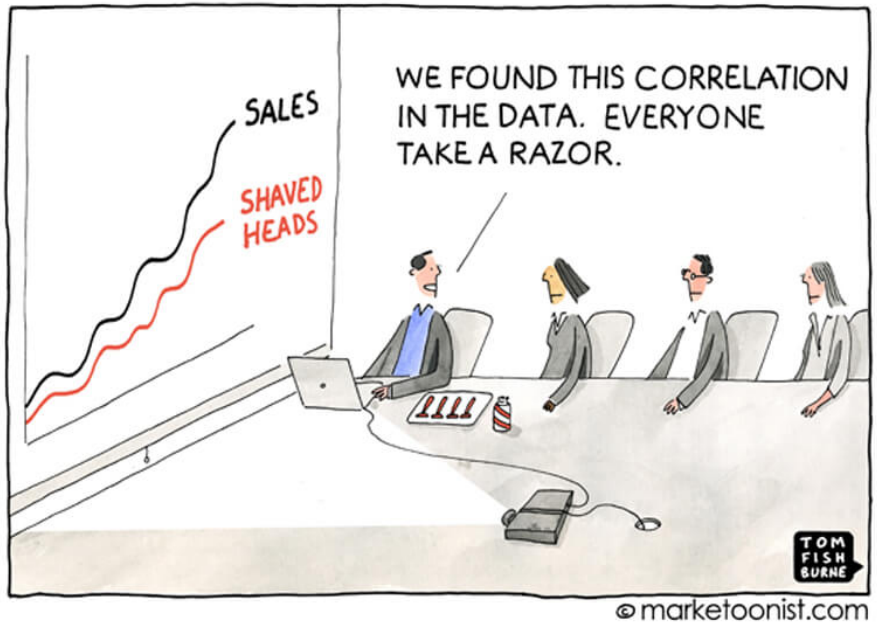
Problems of today's ML - ***Stability***

- Cancer survival rate prediction

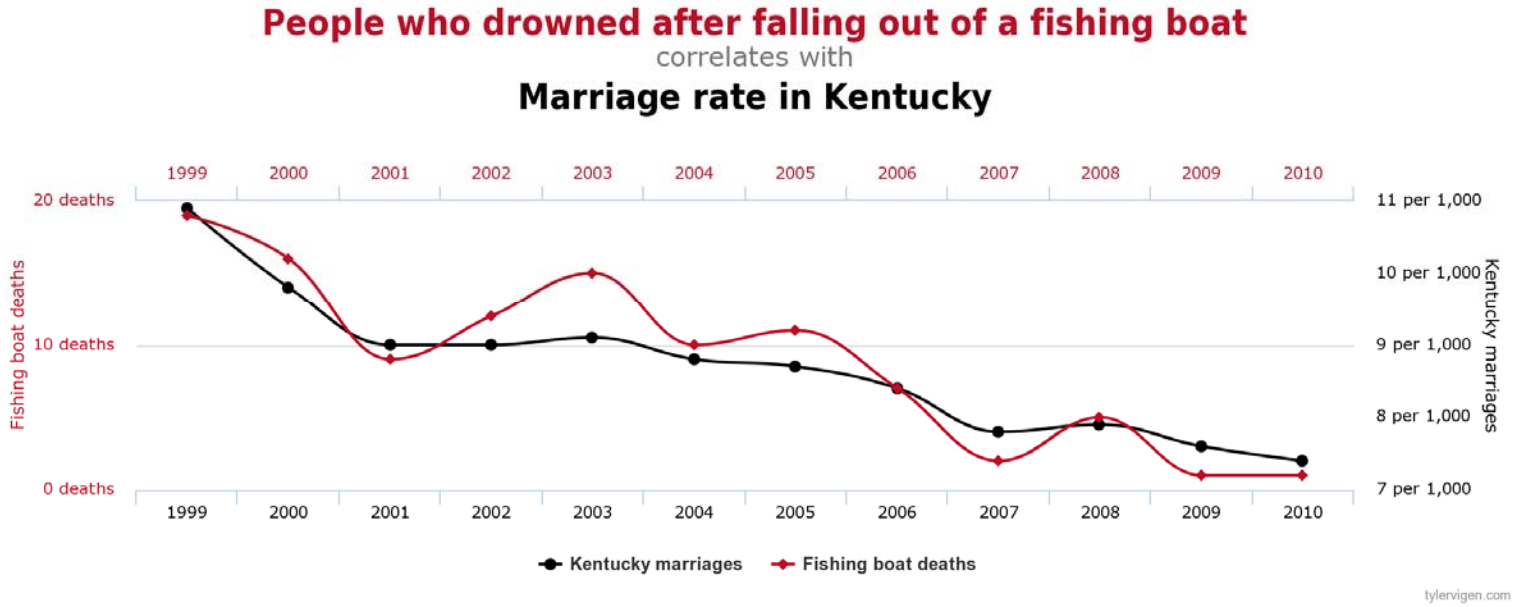


A plausible reason: ***Correlation***

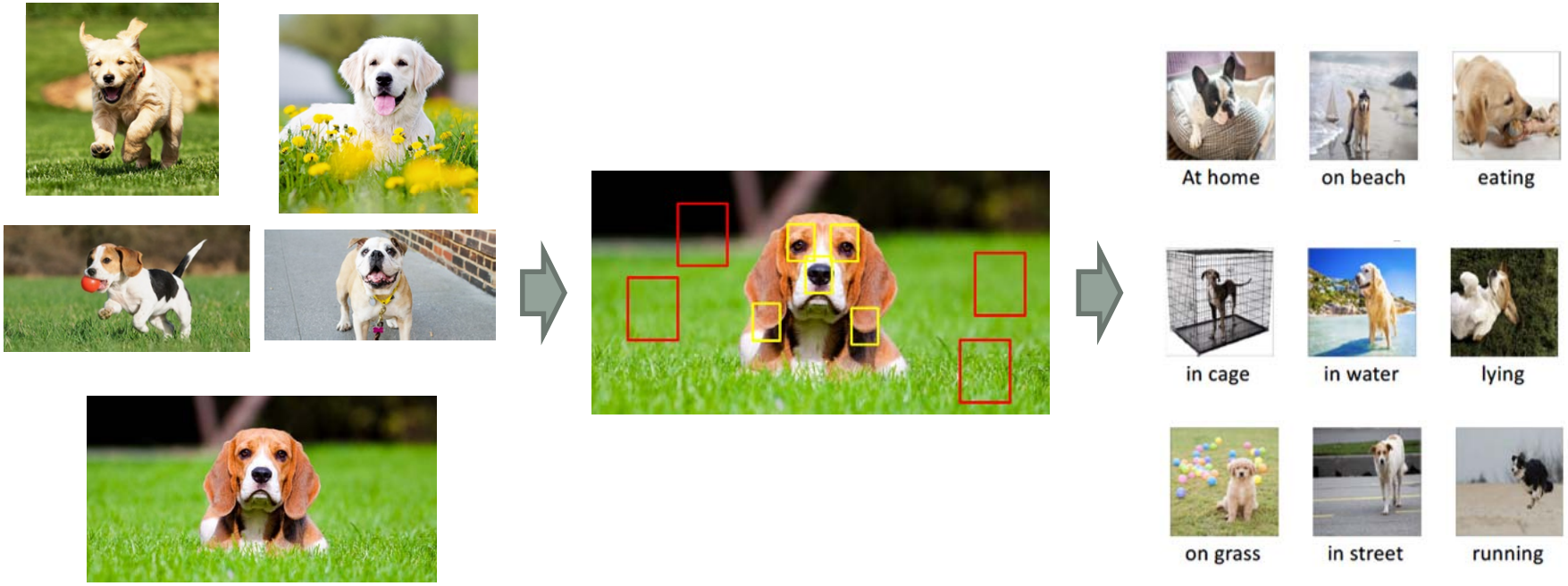
Correlation is the very basics of machine learning.



Correlation is not explainable

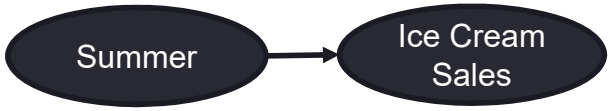


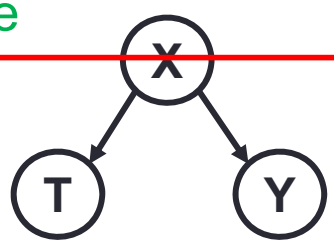
Correlation is '*unstable*'

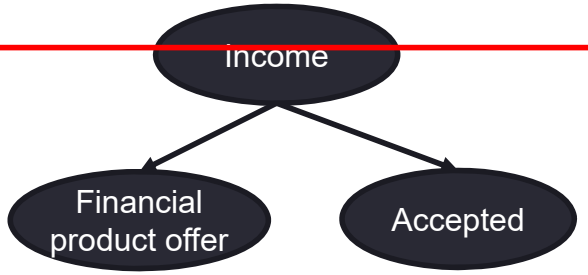


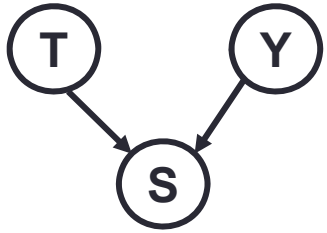
It's not the fault of ***correlation***, but the way we use it

- Three sources of correlation:
  - Causation
    - Causal mechanism
    - Stable and explainable
  - Confounding
    - Ignoring X
    - Spurious Correlation
  - Sample Selection Bias
    - Conditional on S
    - Spurious Correlation







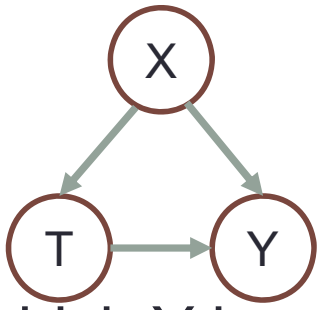


A Practical Definition of Causality

Definition: T causes Y if and only if

changing T leads to a change in Y,

while keeping everything else constant.



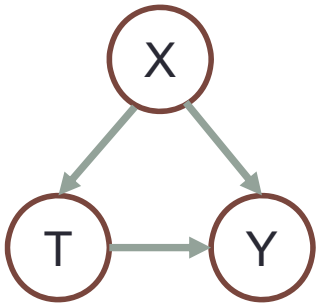
Causal effect is defined as the magnitude by which Y is changed by a unit change in T.

Called the “interventionist” interpretation of causality.

**Interventionist* definition [http://plato.stanford.edu/entries/causation-mani/]

The ***benefits*** of bringing causality into learning

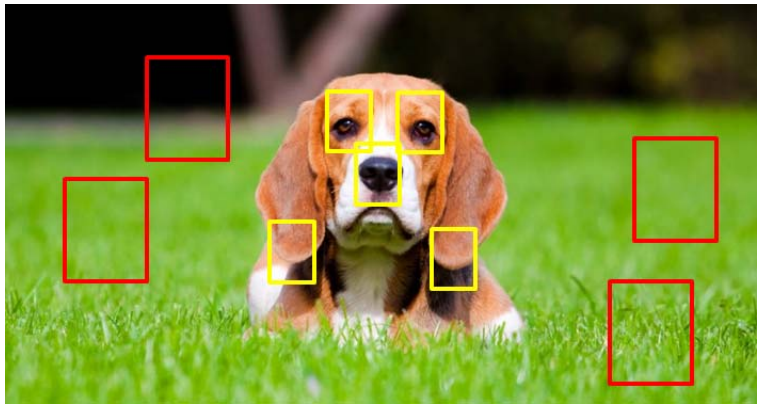
Causal Framework



T：grass
X：dog nose
Y：label

Grass—Label: Strong correlation
Weak causation

Dog nose—Label: Strong correlation
Strong causation



More ***Explainable*** and More ***Stable***

The ***gap*** between causality and learning

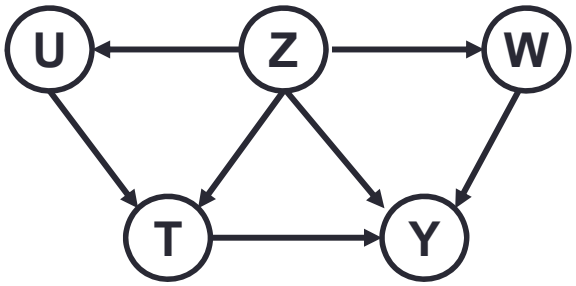
- How to evaluate the outcome?
- Wild environments
  - High-dimensional
  - Highly noisy
  - Little prior knowledge (model specification, confounding structures)
- Targeting problems
  - Understanding v.s. Prediction

How to bridge the gap between *causality* and *(stable) learning*?

Paradigms - Structural Causal Model

A graphical model to describe the causal mechanisms of a system

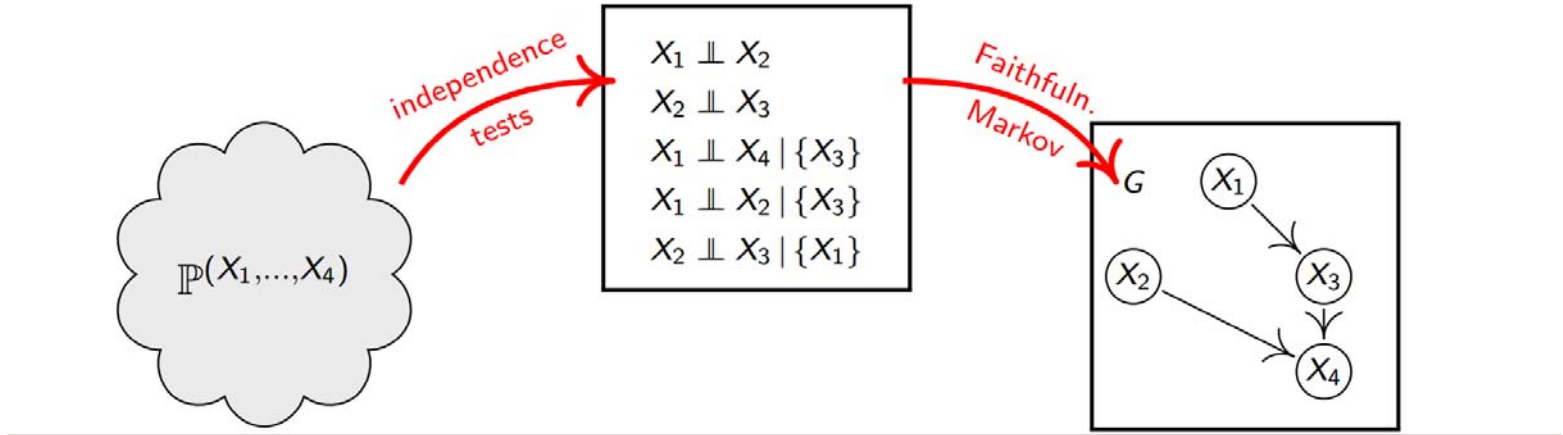
- Causal Identification with back door criterion
- Causal Estimation with do calculus



How to discover the causal structure?

Paradigms – Structural Causal Model

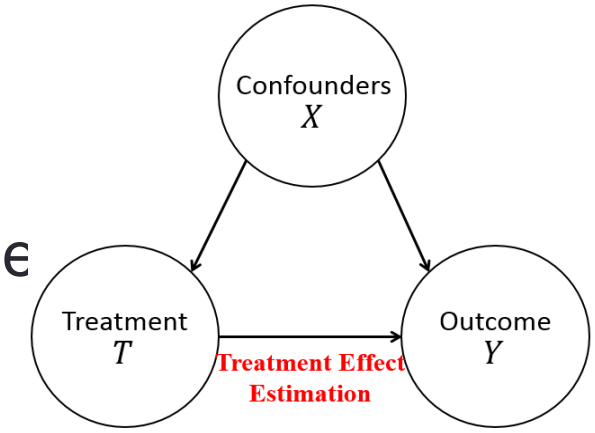
- Causal Discovery
  - Constraint-based: conditional independence
  - Functional causal model based



A ***generative*** model with strong expressive power.
But it induces high complexity.

Paradigms - Potential Outcome Framework

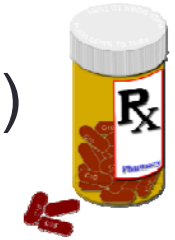
- A simpler setting
  - Suppose the confounders of T are known a priori
- The computational complexity is affordable
  - Under stronger assumptions
  - E.g. all confounders need to be observed



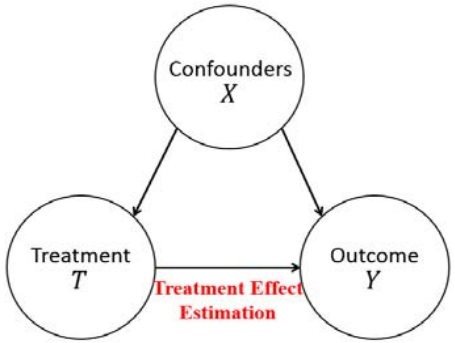
More like a ***discriminative*** way to estimate treatment's partial effect on outcome.

Causal Effect Estimation

- Treatment Variable: $T = 1$ or $T = 0$
- Treated Group $(T = 1)$ and Control Group $(T = 0)$
- Potential Outcome: $Y(T = 1)$ and $Y(T = 0)$
- Average Causal Effect of Treatment (ATE):



$$ATE = E[Y(T = 1) - Y(T = 0)]$$



Counterfactual Problem

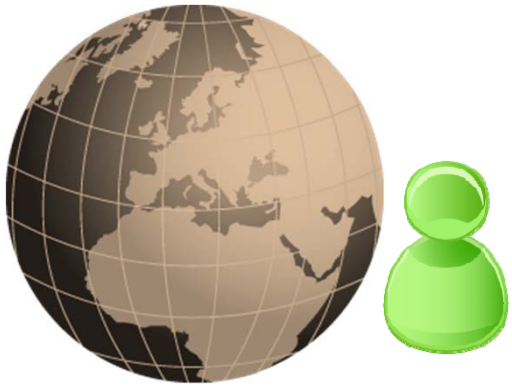
Person	T	$Y_{T=1}$	$Y_{T=0}$
P1	1	0.4	?
P2	0	?	0.6
P3	1	0.3	?
P4	0	?	0.1
P5	1	0.5	?
P6	0	?	0.5
P7	0	?	0.1

- Two key points for causal effect estimation
  - Changing T
  - Keeping everything else constant
- For each person, observe only one: either $Y_{t=1}$ or $Y_{t=0}$
- For different group (T=1 and T=0), something else are not constant

Ideal Solution: Counterfactual World

- Reason about a world that does not exist
- Everything in the counterfactual world is the same as the real world, except the treatment

$Y(T = 1)$



$Y(T = 0)$

Randomized Experiments are the "Gold Standard"

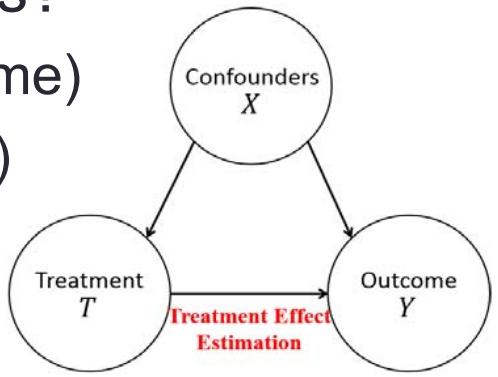
- Drawbacks of Experiments:
  - Cost
  - Unethical
  - Unrealistic

Causal Inference with Observational Data

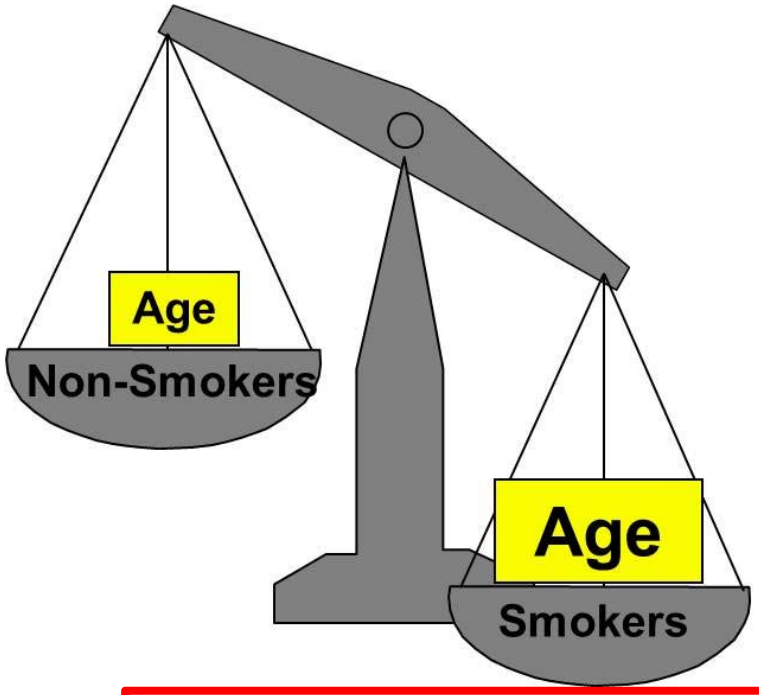
- Counterfactual Problem:

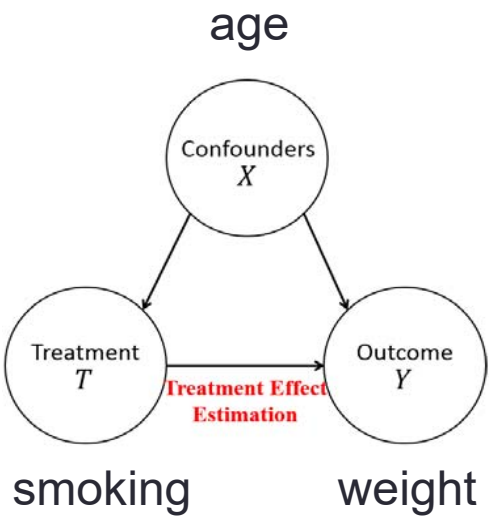
$$Y(T=1) \quad \text{or} \quad Y(T=0)$$

- Can we estimate ATE by directly comparing the average outcome between treated and control groups?
  - Yes with randomized experiments (X are the same)
  - No with observational data (X might be different)



Confounding Effect



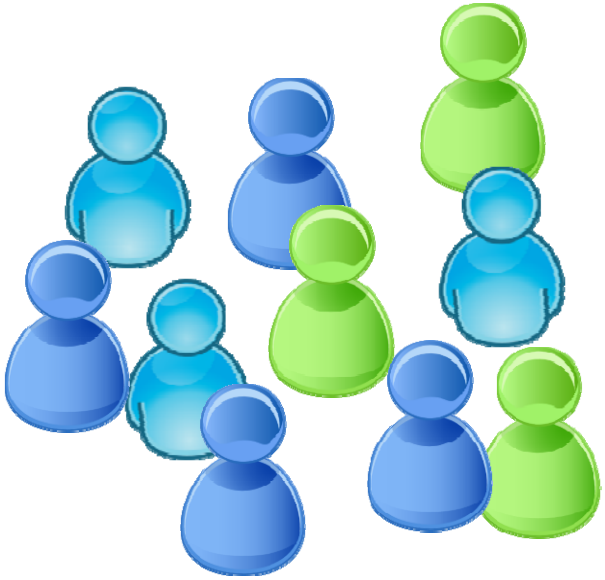


Balancing Confounders' Distribution

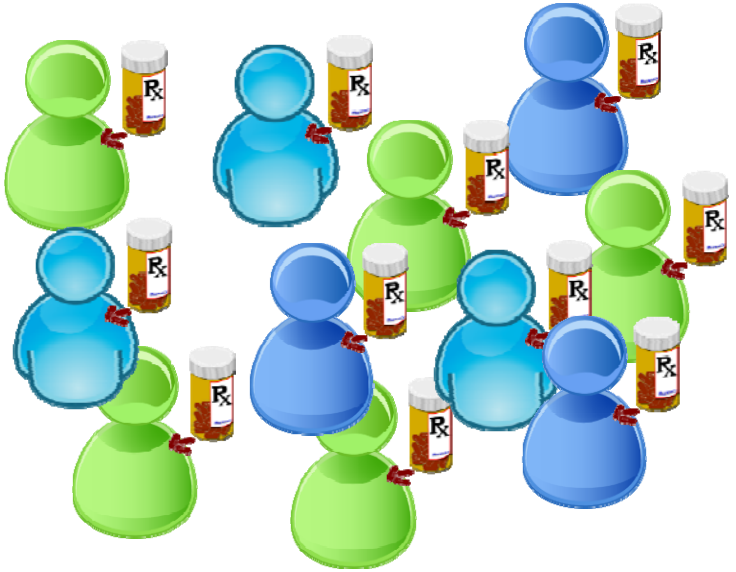
Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

Matching

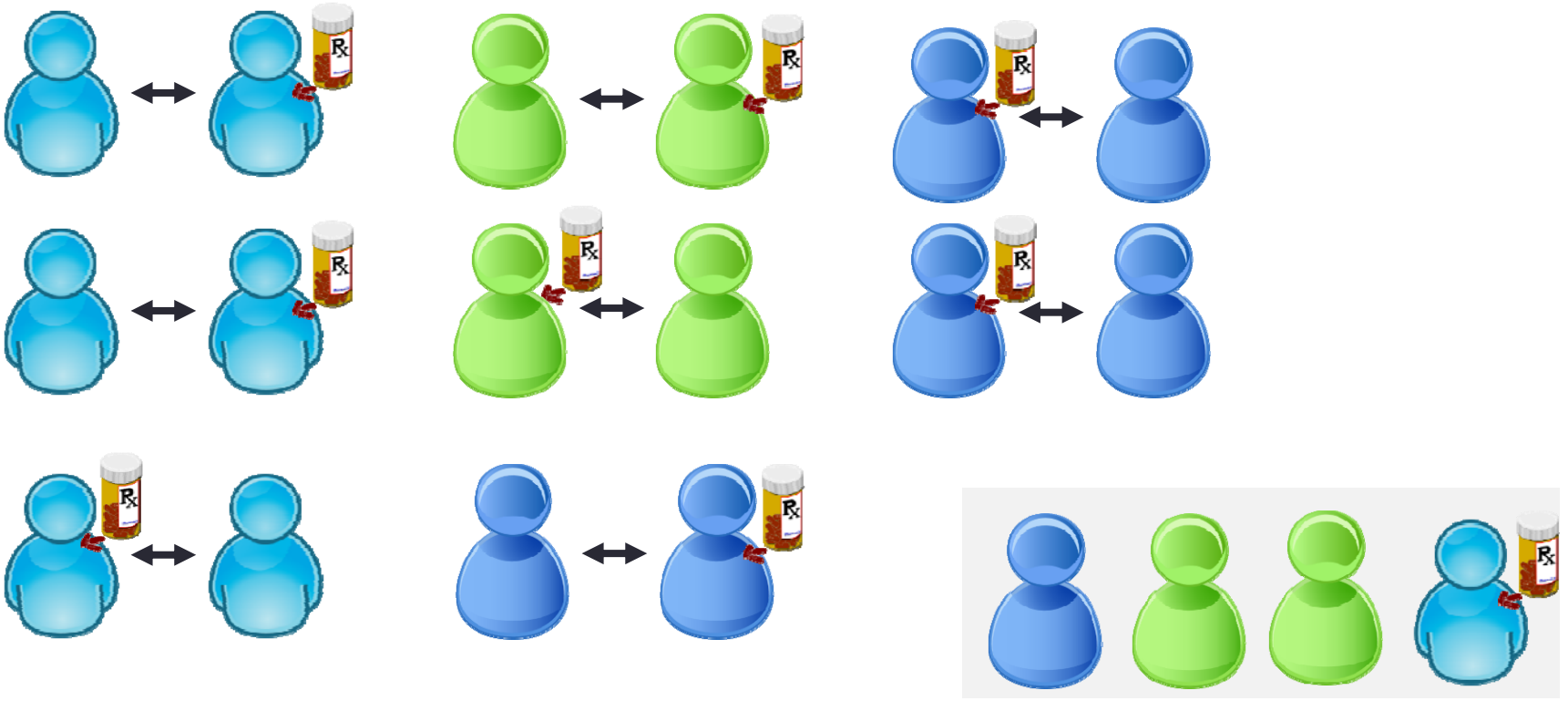


$T = 0$



$T = 1$

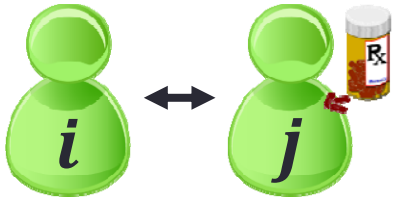
Matching



Matching

- Identify pairs of treated (T=1) and control (T=0) units whose confounders X are similar or even identical to each other

$$Distance(X_i, X_j) \leq \epsilon$$



- Paired units guarantee that the everything else (Confounders) approximate constant
- Small ϵ: less bias, but higher variance
- Fit for low-dimensional settings
- But in high-dimensional settings, there will be few exact matches

Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

Propensity Score Based Methods

- Propensity score $e(X)$ is the probability of a unit to get treated

$$e(X) = P(T = 1|X)$$

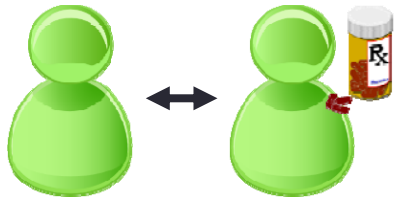
- Then, Donald Rubin shows that the propensity score is sufficient to control or summarize the information of confounders

$$T \perp\!\!\!\perp X \mid e(X) \quad \Rightarrow \quad T \perp\!\!\!\perp (Y(1), Y(0)) \mid e(X)$$

- Propensity scores cannot be observed, need to be estimated

Propensity Score Matching

- Estimating propensity score: $\hat{e}(X) = P(T = 1|X)$
  - **Supervised learning**: predicting a known label T based on observed covariates X.
  - Conventionally, use logistic regression



- Matching pairs by distance between propensity score:

$$Distance(X_i, X_j) \leq \epsilon$$

$$Distance(X_i, X_j) = |\hat{e}(X_i) - \hat{e}(X_j)|$$

- High dimensional challenge: from matching to PS estimation

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3):399–424, 2011.

Inverse of Propensity Weighting (IPW)

- Why weighting with inverse of propensity score?
  - Propensity score induces the distribution bias on confounders X

$$e(X) = P(T = 1|X)$$

Unit	$e(X)$	$1 - e(X)$	#units	#units(T=1)	#units(T=0)
A	0.7	0.3	10	7	3
B	0.6	0.4	50	30	20
C	0.2	0.8	40	8	32

Distribution Bias

Unit	#units(T=1)	#units(T=0)
A	10	10
B	50	50
C	40	40

Confounders are the same!

Reweighting by inverse of propensity score: $w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Inverse of Propensity Weighting (IPW)

- Estimating ATE by IPW [1]:

$$w_i = \frac{T_i}{e_i} + \frac{1 - T_i}{1 - e_i}$$

$$ATE_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{e}(X_i)}$$

  - Interpretation: IPW creates a pseudo-population where the confounders are the same between treated and control groups.
- But requires correct model specification for propensity score
- High variance when e is close to 0 or 1

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Non-parametric solution

- Model specification problem is inevitable
- Can we directly learn sample weights that can balance confounders' distribution between treated and control groups?

Methods for Causal Inference

- **Matching**
- **Propensity Score**
- **Directly Confounder Balancing**

Directly Confounder Balancing

- **Motivation**: The collection of all the moments of variables uniquely determine their distributions.
- **Methods**: Learning sample weights by directly balancing confounders' moments as follows (ATT problem)

$$\min_{W} \| \overline{\mathbf{X}}_t - \mathbf{X}_c^T W \|_2^2$$

The first moments of X on the Treated Group

The first moments of X on the Control Group

With moments, the sample weights can be learned without any model specification.

J. Hainmueller. Entropy balancing for causal effects: A mul- tivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1):25–46, 2012.

Entropy Balancing

$$\begin{aligned} \min_{W} \quad & W \log(W) \\ s.t. \quad & \|\overline{\mathbf{X}}_t - \mathbf{X}_c^T W\|_2^2 = 0 \\ & \sum_{i=1}^{n} W_i = 1, W \succeq 0 \end{aligned}$$

- Directly confounder balancing by sample weights W
- Minimize the entropy of sample weights W

Either know confounders a priori or regard all variables as confounders .
All confounders are balanced equally.

Athey S, et al. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B, 2018, 80(4): 597-623.

Differentiated Confounder Balancing

- **Idea**: Different confounders make different confounding bias
- Simultaneously learn ***confounder weights β*** and ***sample weighs W***.

$$\min \quad \left(\beta^T \cdot (\overline{\mathbf{X}}_t - \mathbf{X}_c^T W)\right)^2$$

- ***Confounder weights*** determine which variable is confounder and its contribution on confounding bias.
- ***Sample weights*** are designed for confounder balancing.

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

Differentiated Confounder Balancing

- General relationship among X, T, and Y:

$$Y = f(\mathbf{X}) + T \cdot g(\mathbf{X}) + \epsilon \Rightarrow \begin{array}{l} ATT = E(g(\mathbf{X}_t)) \\ Y(0) = f(\mathbf{X}) + \epsilon \end{array}$$

$$\begin{aligned} f(\mathbf{X}) &= \mathbf{a}_1\mathbf{X} + \sum_{ij} a_{ij}X_iX_j + \sum_{ijk} a_{ijk}X_iX_jX_k + \cdots + R_n(\mathbf{X}) \\ &= \alpha\mathbf{M}. \qquad \mathbf{M} = (\mathbf{X}, X_iX_j, X_iX_jX_k, \cdots). \end{aligned}$$

Confounder weights

Confounding bias

$$\widehat{ATT} = ATT + \sum_{k=1}^{p} \alpha_k \left(\sum_{i:T_i=1} \frac{1}{n_t} M_{i,k} - \sum_{j:T_j=0} W_j M_{j,k} \right) + \phi(\epsilon).$$

If $\alpha_k = 0$, then M_k is not confounder, no need to balance.
Different confounders have different confounding weights.

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

Differentiated Confounder Balancing

- **Ideas**: simultaneously learn ***confounder weights β*** and ***sample weighs W***.

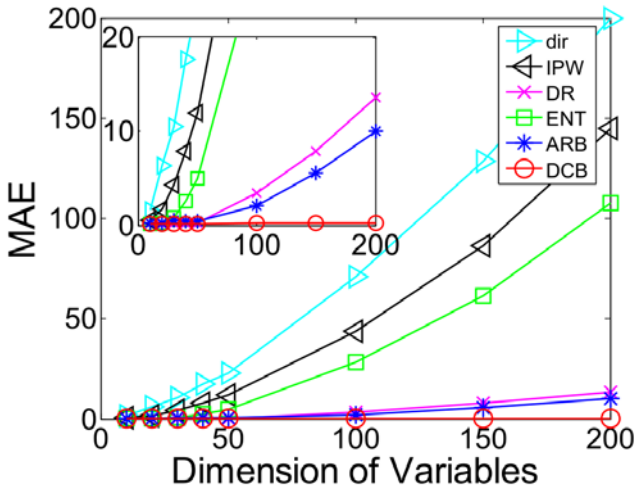
$$\min \quad \left(\beta^T \cdot (\overline{\mathbf{M}}_t - \mathbf{M}_c^T W)\right)^2 + \lambda \sum_{j:T_j=0} (1 + W_j) \cdot (Y_j - M_j \cdot \beta)^2,$$
$$s.t. \quad \|W\|_2^2 \leq \delta, \ \ \|\beta\|_2^2 \leq \mu, \ \ \|\beta\|_1 \leq \nu, \mathbf{1}^T W = 1 \ \ and \ \ W \succeq 0$$

- ***Confounder weights*** determine which variable is confounder and its contribution on confounding bias.
- ***Sample weights*** are designed for confounder balancing.

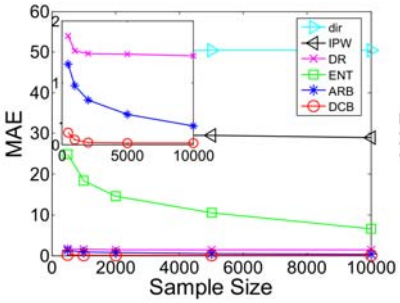
- The ENT algorithm is a special case of DCB algorithm by setting the confounder weights as unit vector.

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

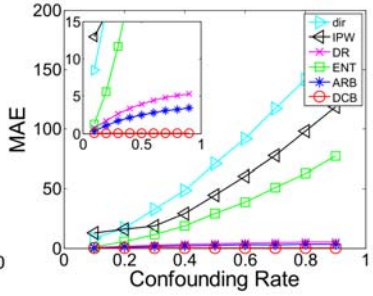
Experiments



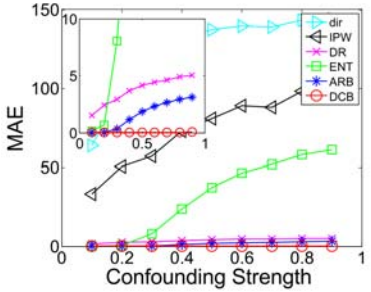
(b) dimension of variables p



(a) sample size n



(c) confounding rate r_c



(d) confounding strength s_c

Variables Set	V-RAW		V-INTERACTION	
Estimator	$\widehat{ATT}$	$Bias$ (SD)	$\widehat{ATT}$	$Bias$ (SD)
$\widehat{ATT}_{dir}$	-8471	10265 (374)	-8471	10265 (374)
$\widehat{ATT}_{IPW}$	-4481	6275 (971)	-4365	6159 (1024)
$\widehat{ATT}_{DR}$	1154	639 (491)	1590	204 (812)
$\widehat{ATT}_{ENT}$	1535	259 (995)	1405	388 (787)
$\widehat{ATT}_{ARB}$	1537	257 (996)	1627	167 (957)
$\widehat{ATT}_{DCB}$	1958	**164** (728)	1836	**43** (716)

LaLonde

Kun Kuang, Peng Cui, et al. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing, KDD 2017, 265–274.

Assumptions of Causal Inference

- **A1: Stable Unit Treatment Value (SUTV):** The effect of treatment on a unit is independent of the treatment assignment of other units

$$P\left(Y_i \middle| T_i, T_j, X_i\right) = P(Y_i | T_i, X_i)$$

- **A2: Unconfounderness:** The distribution of treatment is independent of potential outcome when given the observed variables

$$T \perp \left(Y(0), Y(1)\right) | \, X$$

No unmeasured confounders

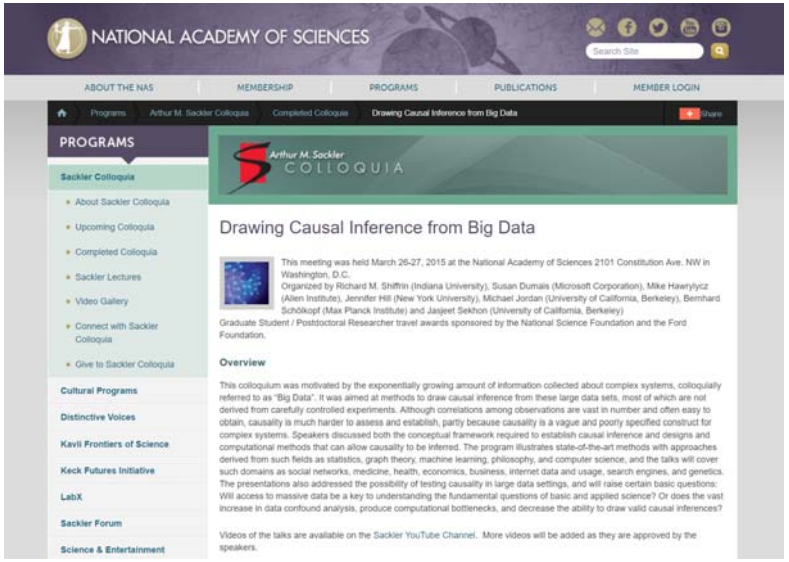
- **A3: Overlap:** Each unit has nonzero probability to receive either treatment status when given the observed variables

$$0 < P(T = 1 | X = x) < 1$$

Sectional Summary

- Progress has been made to draw causality from big data.
- From single to group
- From binary to continuous
- Weak assumptions

Ready for Learning?



Outline

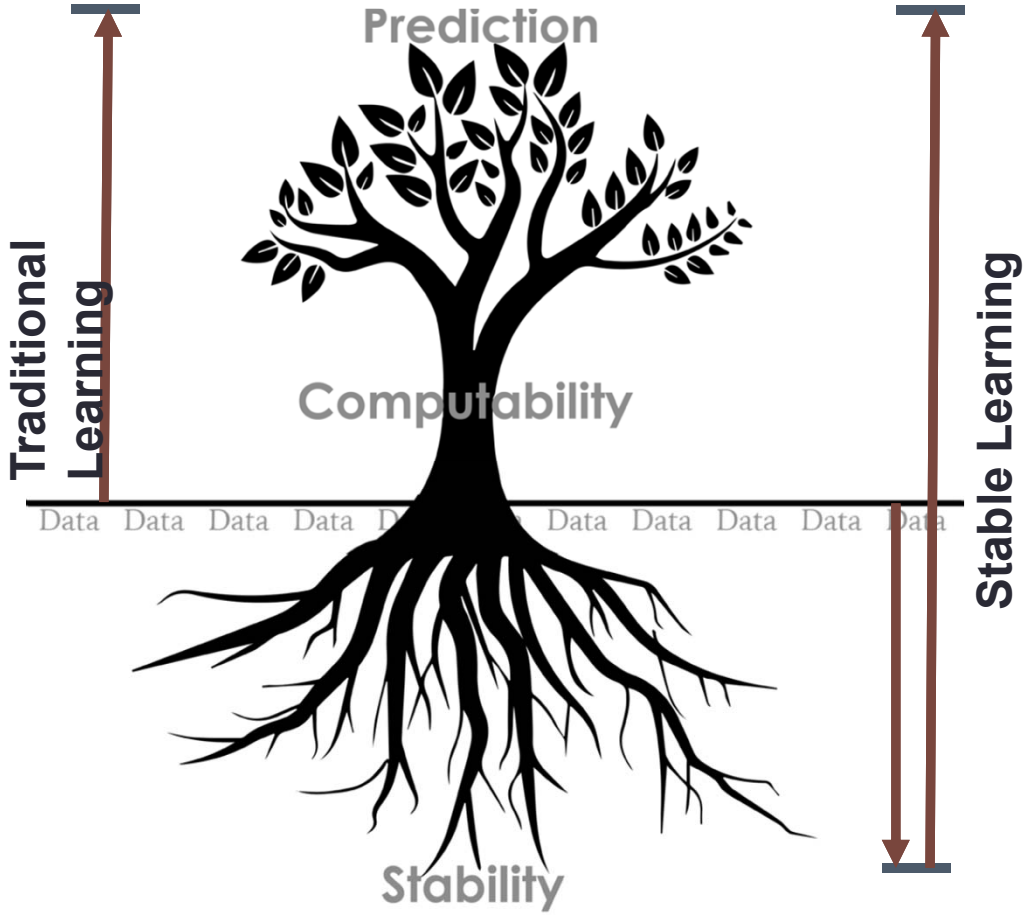
- Motivation of human-like learning
- Learning to learn (association)
- Causal inference
- **Stable learning**
- The NICO dataset

Stability and Prediction

Prediction Performance

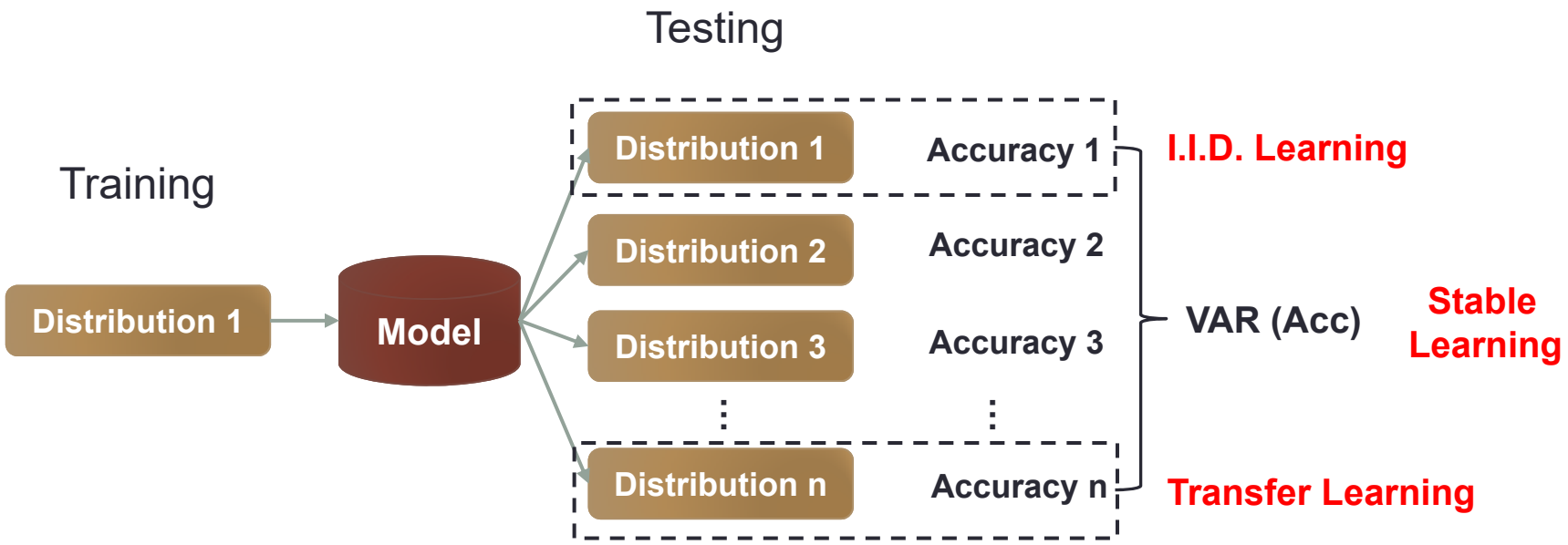
Learning Process

True Model



Bin Yu (2016), Three Principles of Data Science: predictability, computability, stability

Stable Learning



Stability and Robustness

- Robustness
  - More on prediction performance over data perturbations
  - ***Prediction*** performance-driven
- Stability
  - More on the true model
  - Lay more emphasis on ***Bias***
  - Sufficient for robustness

Stable learning is a (intrinsic?) way to realize robust prediction

Stability

- Statistical stability holds if statistical conclusions are robust to appropriate perturbations to data.
  - Prediction Stability
  - Estimation Stability

Bernoulli **19**(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

Departments of Statistics and EECS, University of California at Berkeley, Berkeley, CA 94720, USA.
E-mail: binyu@stat.berkeley.edu

Prediction Stability

- Lasso

$$\hat{\beta}(\lambda) = \underset{\beta \in R^p}{\arg} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

- Prediction Stability by Cross-validation
  - n data units are randomly partitioned into V blocks, each block has d = [n/V] units.
  - Leave one out: training on (n-d) units, validating on d units.
  - CV does not provide a good interpretable model because Lasso+CV is unstable.

Estimation Stability

- Estimation Stability:
  - Mean regression function:
  - Variance of function m:
  - Estimation Stability:

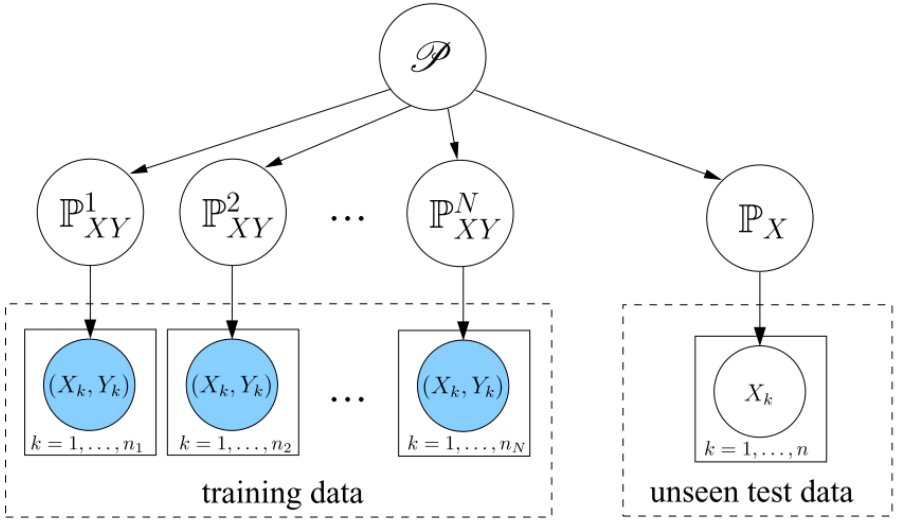
$$\hat{m}(\tau) = \frac{1}{V}\sum_{v} X\hat{\beta}_v(\tau),$$

$$\hat{T}(\tau) = \frac{n-d}{d}\frac{1}{V}\sum_{v}\left(\left\|X\hat{\beta}_v(\tau) - \hat{m}(\tau)\right\|^2\right).$$

$$ES(\tau) = \frac{1/V\sum_v \|X\hat{\beta}_v(\tau) - \hat{m}(\tau)\|^2}{\hat{m}^2(\tau)} = \frac{d}{n-d}\frac{\hat{T}(\tau)}{\hat{m}^2(\tau)}$$

ES+CV is better than Lasso+CV

Domain Generalization / Invariant Learning



- Given data from different observed environments $e \in \mathcal{E}$.
  $(X^e, Y^e) \sim F^e, \quad e \in \mathcal{E}$
- The task is to predict Y given X such that the prediction works well (is “robust”) for “all possible” (including unseen) environments

Domain Generalization

- **Assumption**: the conditional probability P(Y|X) is stable or invariant across different environments.
- **Idea**: taking knowledge acquired from a number of related domains and applying it to previously unseen domains
- **Theorem**: Under reasonable technical assumptions. Then with probability at least

$$1-\delta$$

$$\sup_{\|f\|_{\mathcal{H}}\leq 1}\left|\mathbb{E}^*_{\mathscr{P}}\mathbb{E}_{\mathbb{P}}\ell(f(\tilde{X}_{ij}),Y_i)-\mathbb{E}_{\hat{\mathbb{P}}}\ell(f(\tilde{X}_{ij}),Y_i)\right|^2$$

$$\leq c_1\cdot\underbrace{\mathbb{V}_{\mathcal{H}}(\mathbb{P}^1,\mathbb{P}^2,\ldots,\mathbb{P}^N)}_{\textit{distributional variance}}+\underbrace{c_2\frac{N\cdot(\log\delta^{-1}+2\log N)}{n}+c_3\frac{\log\delta^{-1}}{N}+\frac{c_4}{N}}_{\textit{vanish as } N,n\to\infty}$$

Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature. ICML 2013.

Invariant Prediction

- **Invariant Assumption**: There exists a subset $S \in X$ is causal for the prediction of Y. and the conditional distributio for all $e \in \mathcal{E}$, X^e has an arbitrary distribution and .

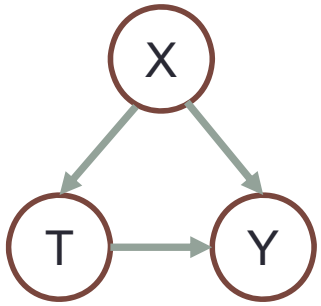
$$Y^e = g(X^e_{S^*}, \varepsilon^e), \qquad \varepsilon^e \sim F_\varepsilon \text{ and } \varepsilon^e \perp\!\!\!\perp X^e_{S^*}$$

$$Y^e \leftarrow \sum_{k \in \mathrm{pa}(Y)} \underbrace{\beta_{Y,k}}_{\forall e} X^e_k + \underbrace{\varepsilon^e_Y}_{\sim F_\varepsilon \forall e \in \mathcal{G}}$$

- **Idea: Linking to causality**
  - Structural Causal Model (Pearl 2009):
  - The parent variables of Y in SCM satisfies Invariant Assumption
  - The causal variables lead to invariance w.r.t. “all” possible environments

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016*

From *Variable Selection* to *Sample Reweighting*



Typical Causal Framework

Directly Confounder Balancing

Given a feature T

↓

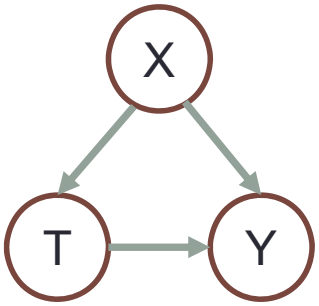
Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Sample reweighting can make a variable independent of other variables.

Global Balancing: Decorrelating Variables



Typical Causal Framework

Global Balancing

Given **ANY** feature T

↓

Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Partial effect can be regarded as causal effect. Predicting with causal variables is stable across different environments.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

Theoretical Guarantee

PROPOSITION 3.3. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^T \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2, \quad (4)$$

↓

0

PROOF. Since $\|\cdot\| \geq 0$, Eq. (8) can be simplified to $\forall j, \forall k \neq j$

$$\lim_{n\to\infty} \left(\frac{\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=1} W_i}{\sum_{i:\mathbf{X}_{i,j}=1} W_i} - \frac{\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=0} W_i}{\sum_{i:\mathbf{X}_{i,j}=0} W_i} \right) = 0$$

with probability 1. For W^*, from Lemma 3.1, $0 < P(\mathbf{X}_i = x) < 1$, $\forall x, \forall i, t = 1$ or 0,

$$\begin{aligned} \lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{i:\mathbf{X}_{i,j}=t} W_i^* &= \lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{x:x_j=t} \sum_{i:\mathbf{X}_i=x} W_i^* \\ &= \lim_{n\to\infty} \textstyle\sum_{x:x_j=t} \tfrac{1}{n} \sum_{i:\mathbf{X}_i=x} \tfrac{1}{P(\mathbf{X}_i=x)} \\ &= \lim_{n\to\infty} \textstyle\sum_{x:x_j=t} P(\mathbf{X}_I = x) \cdot \tfrac{1}{P(\mathbf{X}_i=x)} = 2^{p-1} \end{aligned}$$

with probability 1 (Law of Large Number). Since features are binary,

$$\lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=1} W_i^* = 2^{p-2}$$

$$\lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{i:\mathbf{X}_{i,j}=0} W_i^* = 2^{p-1}, \quad \lim_{n\to\infty} \tfrac{1}{n} \textstyle\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=0} W_i^* = 2^{p-2}$$

and therefore, we have following equation with probability 1:

$$\lim_{n\to\infty} \left(\frac{\mathbf{X}_{\cdot,k}^T (W^* \odot \mathbf{X}_{\cdot,j})}{W^{*T} \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,k}^T (W^* \odot (1-\mathbf{X}_{\cdot,j}))}{W^{*T} (1-\mathbf{X}_{\cdot,j})} \right) = \frac{2^{p-2}}{2^{p-1}} - \frac{2^{p-2}}{2^{p-1}} = 0.$$

□

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

Causal Regularizer

Set feature *j* as treatment variable

$$\sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2,$$

All features excluding treatment *j*

Sample Weights

Indicator of treatment status

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM**, 2018*.

Causally Regularized Logistic Regression

$$\min \quad \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (x_i \beta))),$$

$$s.t. \quad \sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2 \leq \lambda_1,$$

$$W \geq 0, \quad \|W\|_2^2 \leq \lambda_2, \quad \|\beta\|_2^2 \leq \lambda_3, \quad \|\beta\|_1 \leq \lambda_4,$$

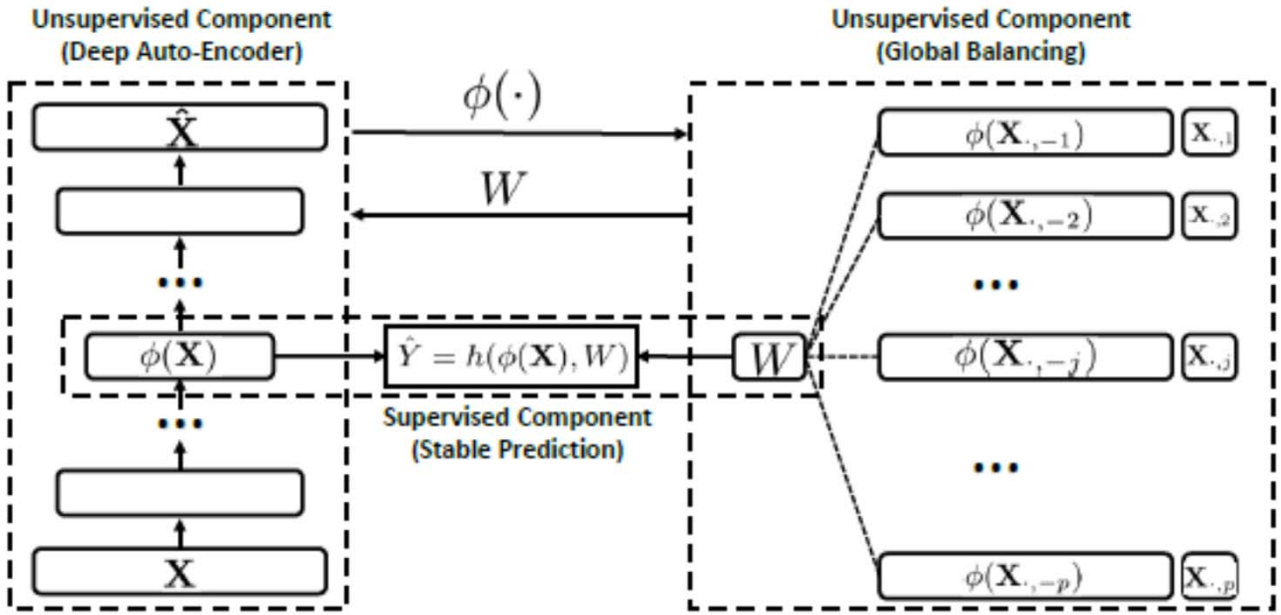
$$(\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_5,$$

Sample reweighted logistic loss

Causal Contribution

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM**, 2018*.

From Shallow to Deep - DGBR

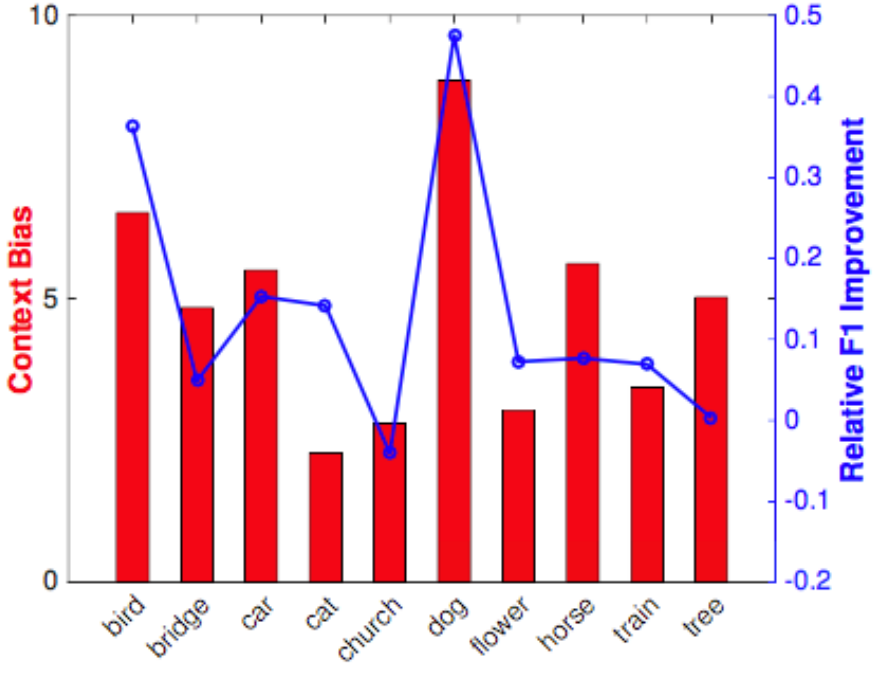


Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

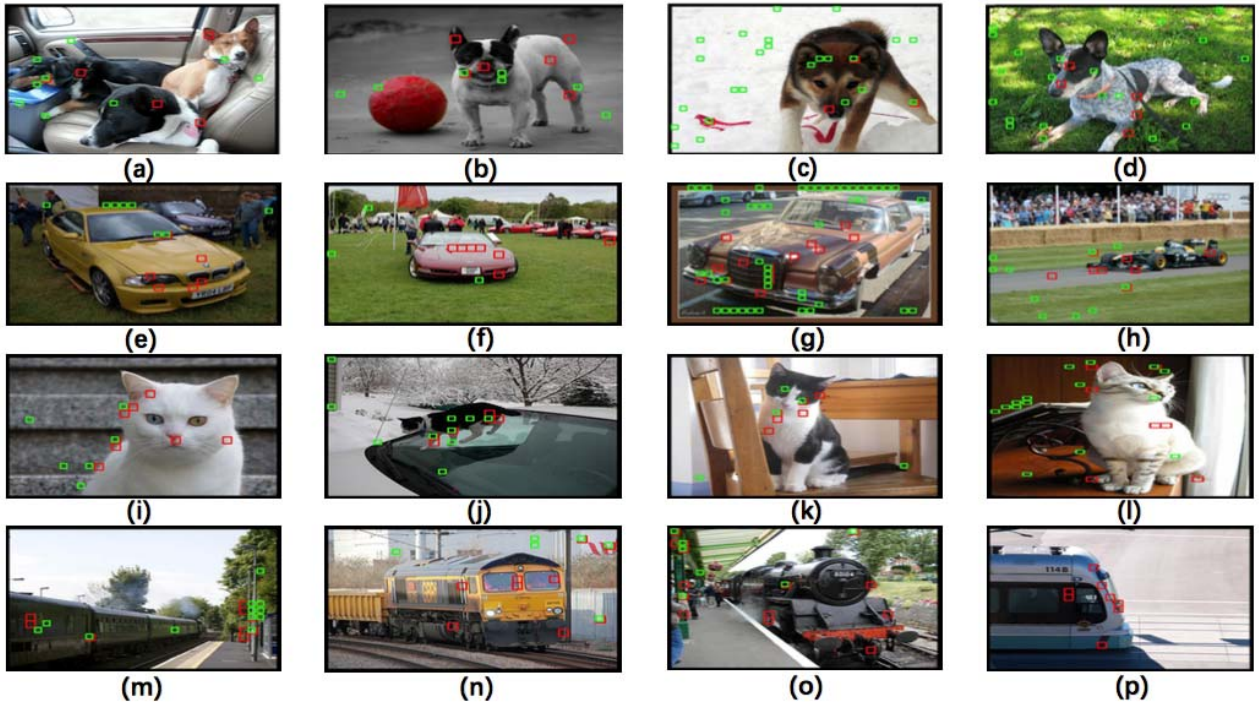
Experiment 1 – non-i.i.d. image classification

- Source: ***YFCC100M***
- Type: high-resolution and multi-tags
- Scale: 10-category, each with nearly 1000 images
- Method: select 5 ***context tags*** which are frequently co-occurred with the ***major tag*** (category label)

Experimental Result - insights



Experimental Result - insights



Experiment 2 – online advertising

- Environments generating:
  - Separate the whole dataset into 4 environments by users' age, including $Age \in [20,30)$, $Age \in [30,40)$, $Age \in [40,50)$, and $Age \in [50,100)$.

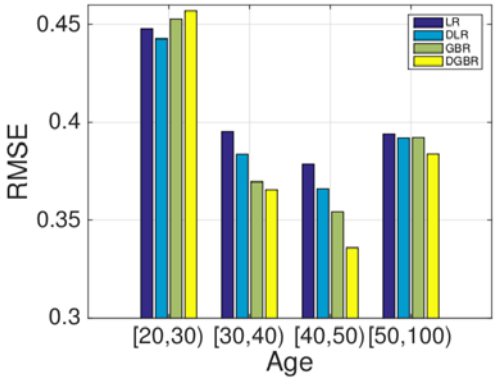


Fig. 15: Prediction across environments separated by age. The models are trained on dataset where uses' $Age \in [20, 30)$, but tested on various datasets with different users' age range.

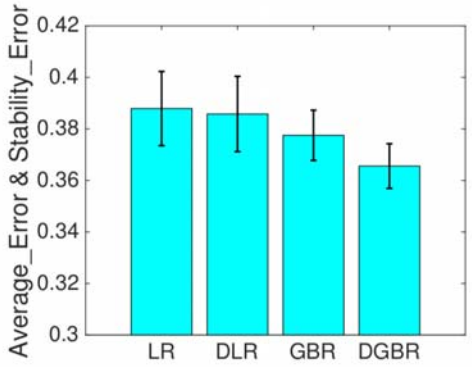
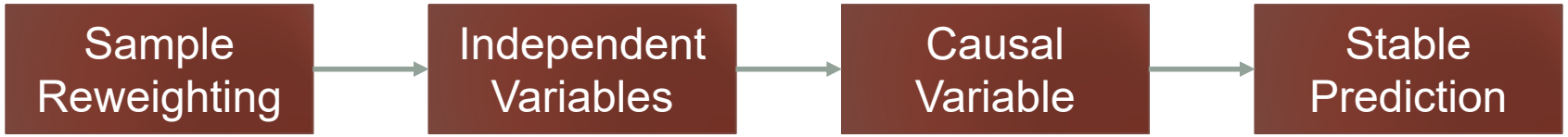


Fig. 16: $Average\_Error$ and $Stability\_Error$ of all algorithms across environments after fixing $P(Y)$ as the same with its value on global dataset.

From ***Causal*** problem to ***Learning*** problem

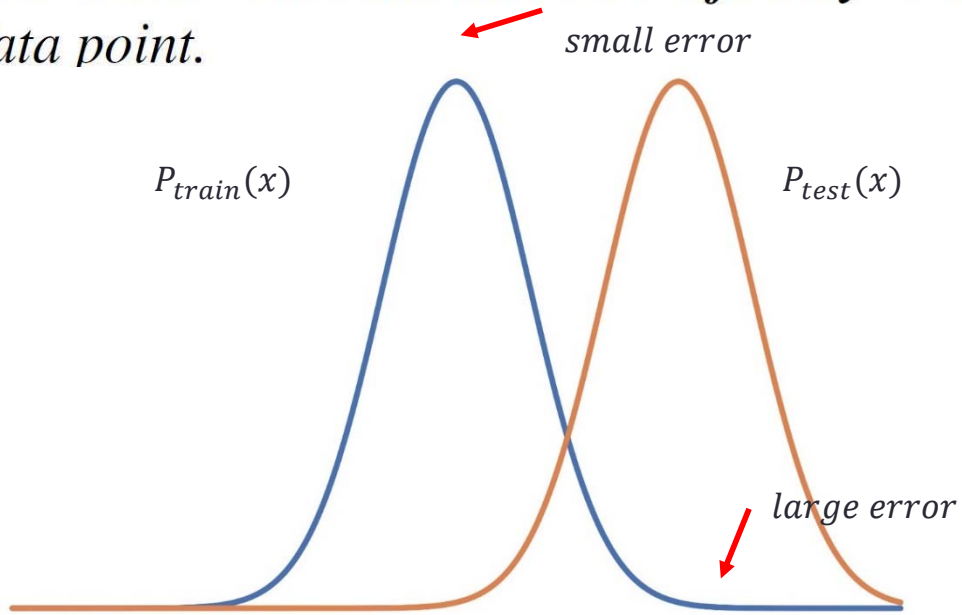
- Previous logic:



- More direct logic:

Thinking from the ***Learning*** end

Problem 1. ***(Stable Learning)*** *: Given the target y and p input variables $x = [x_1, \ldots, x_p] \in \mathbb{R}^p$, the task is to learn a predictive model which can achieve* ***uniformly*** *small error on* ***any*** *data point.*



Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Stable Learning of Linear Models

- Consider the linear regression with misspecification bias

$$y = x^\top \overline{\beta}_{1:p} + \overline{\beta}_0 + b(x) + \epsilon$$

Goes to infinity when perfect collinearity exists!

Bias term with bound $b(x) \leq \delta$

- By accurately estimating $\overline{\beta}$ with the property that $b(x)$ is uniformly small for all x, we can achieve stable learning.
- However, the estimation error caused by misspecification term can be as bad as $\|\hat{\beta} - \overline{\beta}\|_2 \leq 2(\delta/\gamma) + \delta$, where γ^2 is the smallest eigenvalue of centered covariance matrix.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Toy Example

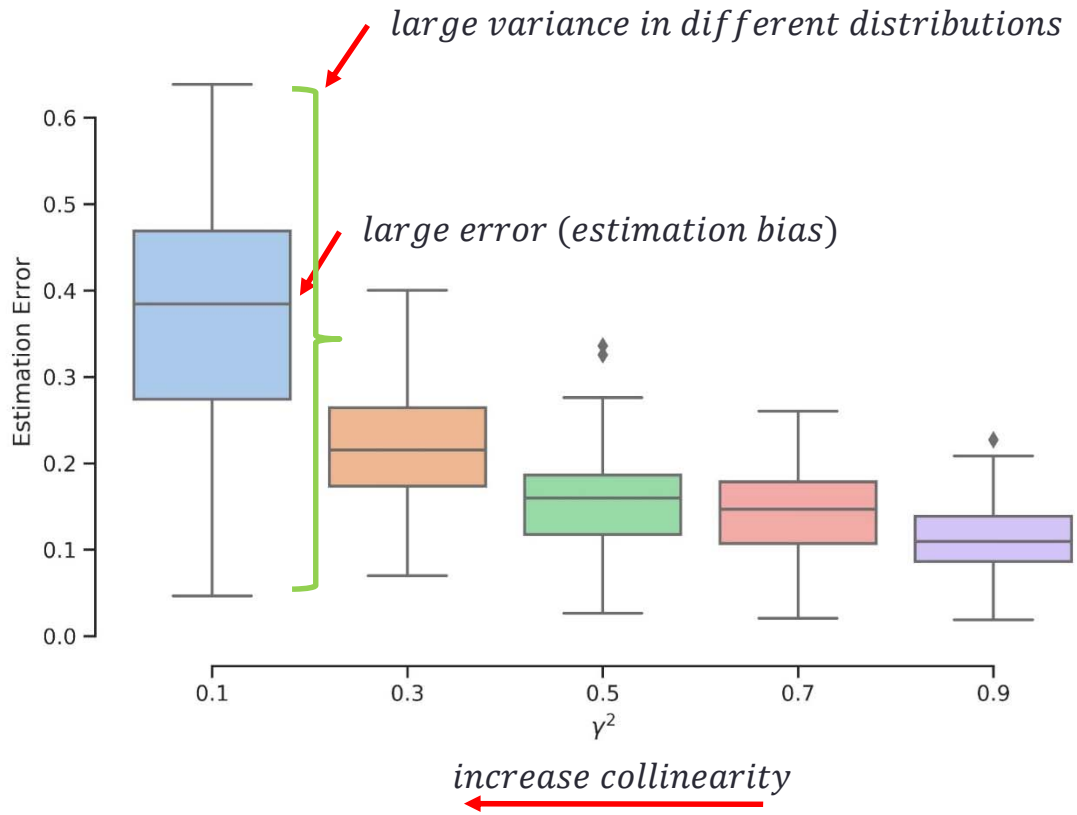
- Assume the design matrix X consists of two variables X_1, X_2, generated from a multivariate normal distribution:

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- By changing ρ, we can simulate different extent of collinearity.
- To induce bias related to collinearity, we generate bias term $b(X)$ with $b(X) = Xv$, where v is the eigenvector of centered covariance matrix corresponding to its smallest eigenvalue γ^2.
- The bias term is sensitive to collinearity.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Simulation Results



Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Reducing collinearity by sample reweighting

Idea: Learn a new set of ***sample weights*** $w(x)$ to decorrelate the input variables and increase the smallest eigenvalue

- Weighted Least Square Estimation

$$\hat{\beta} = \arg\min_{\beta} \mathbf{E}_{(x)\sim D} w(x) \left(x^\top \beta_{1:p} + \beta_0 - y\right)^2$$

which is equivalent to

$$\hat{\beta} = \arg\min_{\beta} \mathbf{E}_{(x)\sim \tilde{D}} \left(x^\top \beta_{1:p} + \beta_0 - y\right)^2$$

So, how to find an "oracle" distribution $\tilde{D}$ which holds the desired property?

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Sample Reweighted Decorrelation Operator

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \xrightarrow{\text{Decorrelation}} \tilde{\mathbf{X}} = \begin{pmatrix} x_{i1} & \dots & x_{rl} & \dots \\ x_{j1} & \dots & x_{sl} & \dots \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & \dots & x_{tl} & \dots \end{pmatrix}$$

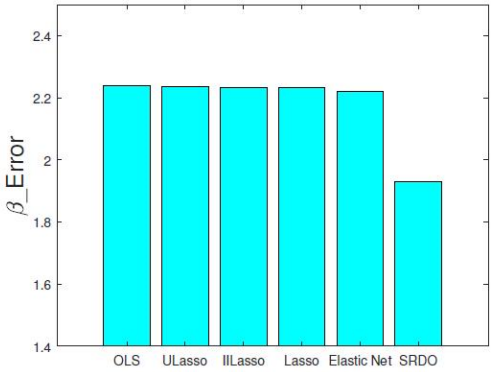
where i, j, k, r, s, t are drawn from $1 \dots n$ at random

- By treating the different columns independently while performing random resampling, we can obtain a column-decorrelated design matrix with the same marginal as before.
- Then we can use density ratio estimation to get $w(x)$.

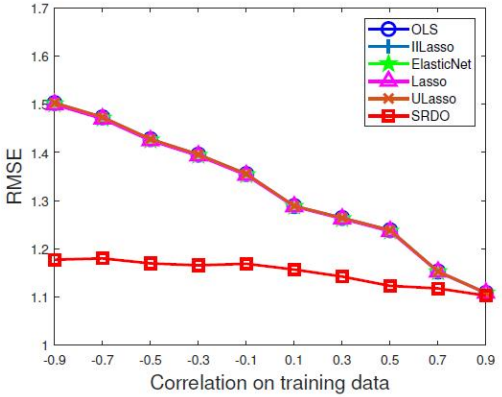
Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

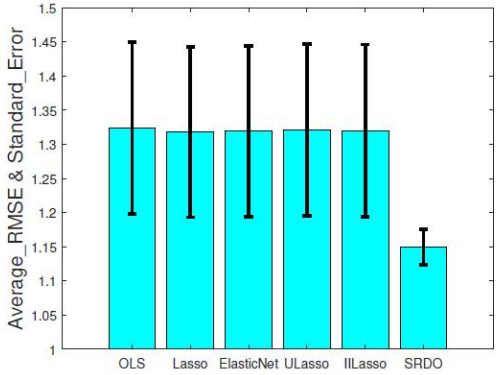
Experimental Results

- Simulation Study



(a) Estimation error





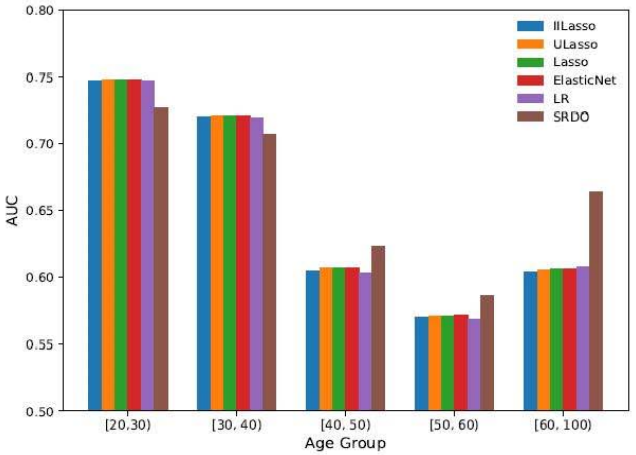
(b) Prediction error over different test environments

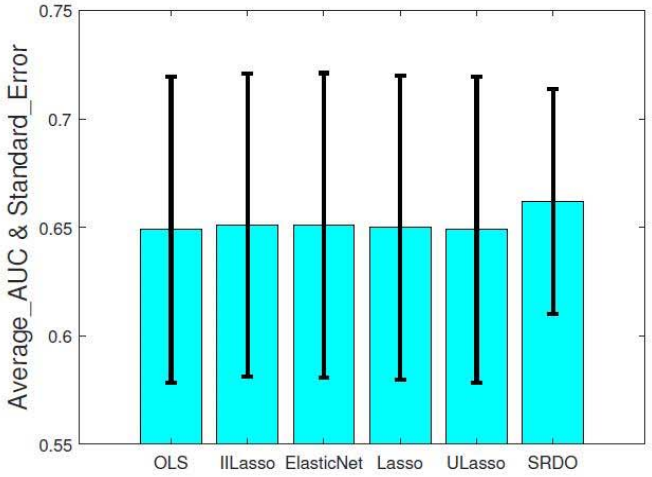
(c) Average prediction error&stability

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Experimental Results

- Regression
- Classification





(a) AUC over different test environments. (b) Average AUC of all the environments and stability.

Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Disentanglement Representation Learning

From decorrelating input variables to learning disentangled representation

- Learning Multiple Levels of Abstraction
  - The big payoff of deep learning is to allow learning higher levels of abstraction
  - Higher-level abstractions disentangle the factor of variation, which allows much easier generalization and transfer

Yoshua Bengio, From Deep Learning of Disentangled Representations to Higher-level Cognition. (2019). YouTube. Retrieved 22 February 2019.

Disentanglement for Causality

- Causal / mechanism independence
  - Independently Controllable Factors *(Thomas, Bengio et al., 2017)*

selectively change

A policy π_k

correspond to value

A representation f_k

$$sel(s,a,k) = \mathbb{E}_{s' \sim \mathcal{P}^a_{ss'}} \left[\frac{|f_k(s') - f_k(s)|}{\sum_{k'} |f_{k'}(s') - f_{k'}(s)|} \right]$$

- Opti $\underbrace{\mathbb{E}_s[\frac{1}{2}||s - g(f(s))||_2^2]}_{\mathcal{L}_{ae} \text{ the reconstruction error}} - \lambda \underbrace{\sum_k \mathbb{E}_s[\sum_a \pi_k(a|s) sel(s,a,k)]}_{\mathcal{L}_{sel} \text{ the disentanglement objective}}.$

Require subtle design on the policy set to guarantee causality.

Sectional Summary

- Causal inference provide valuable insights for stable learning
- Complete causal structure means data generation process, necessarily leading to stable prediction
- Stable learning can also help to advance causal inference
- Performance driven and practical applications

Benchmark is important!

Outline

- Motivation of human-like learning
- Learning to learn (association)
- Causal inference
- Stable learning
- **The NICO dataset**

Non-I.I.D. Image Classification

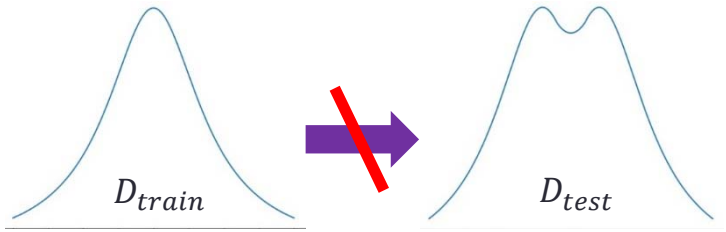
- Non I.I.D. Image Classification

$$\psi(D_{train} = (X_{train}, Y_{train})) \neq \psi(D_{test} = (X_{test}, Y_{test}))$$

- Two tasks
  - Targeted Non-I.I.D. Image Classification
    - Have prior knowledge on testing data
    - e.g. transfer learning, domain adaptation
  - General Non-I.I.D. Image Classification
    - Testing is unknown, no prior
    - more practical & realistic

known

unknown



Existence of Non-I.I.Dness

- One metric (NI) for Non-I.I.Dness

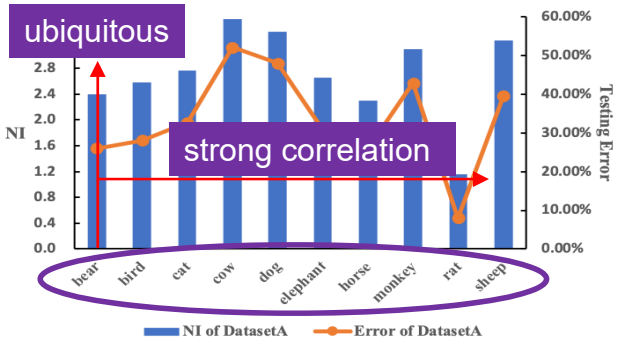
Definition 1 ***Non-I.I.D. Index (NI)*** *Given a feature extractor* $g_\varphi(\cdot)$ *and a class* C, ***the degree of distribution shift*** *between training data* D^C_{train} *and testing data* D^C_{test} *is defined as:*

$$NI(C) = \left\| \frac{\overline{g_\varphi(X^C_{train})} - \overline{g_\varphi(X^C_{test})}}{\sigma(g_\varphi(X^C_{train} \cup X^C_{test}))} \right\|_2,$$

Distribution shift

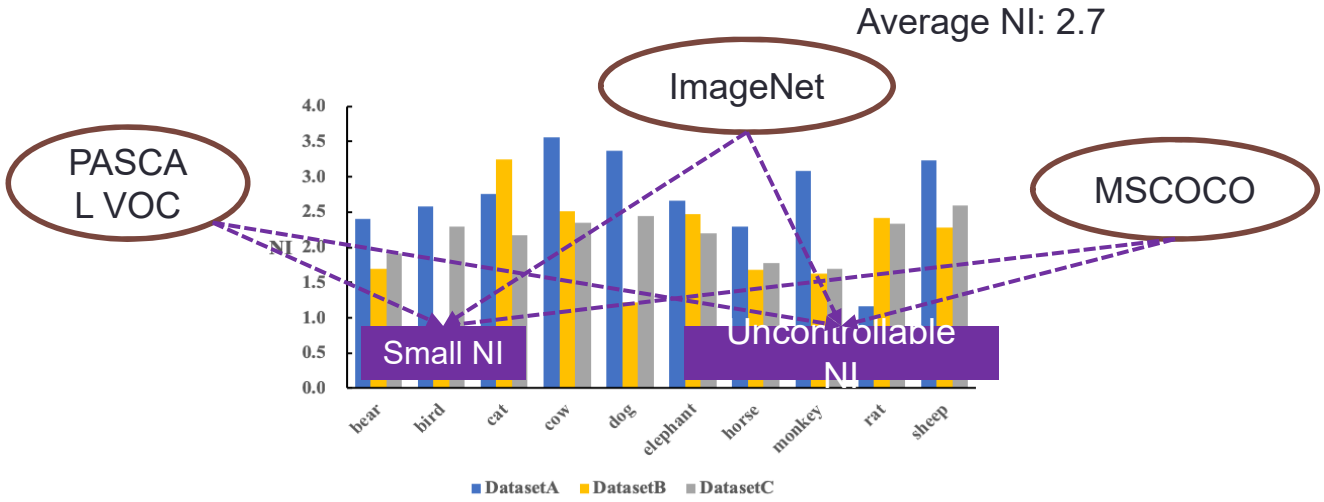
For normalization

- Existence of Non-I.I.Dness on Dataset consisted of 10 subclasses from ImageNet
- For each class
  - Training data
  - Testing data
  - CNN for prediction



Related Datasets

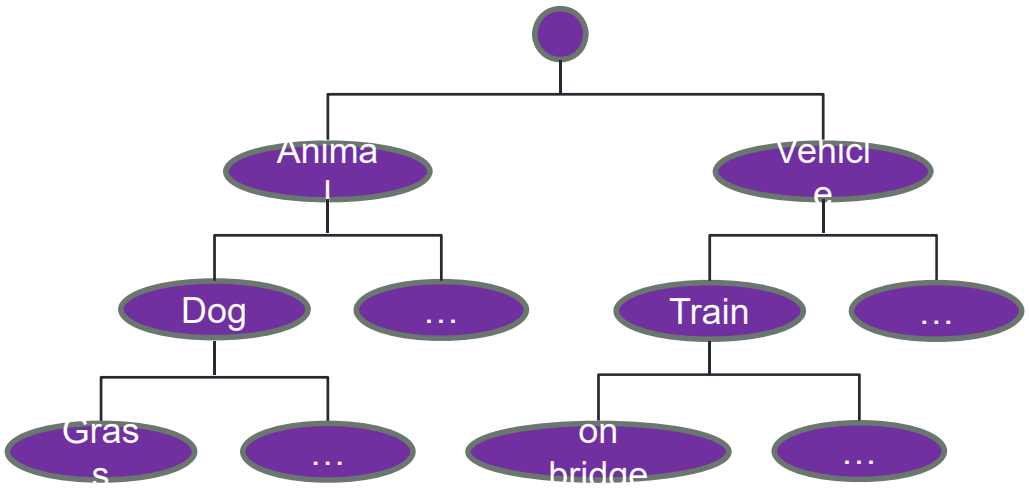
- DatasetA & DatasetB & DatasetC
  - NI is ubiquitous, but small on these datasets
  - NI is Uncontrollable, not friendly for Non IID setting

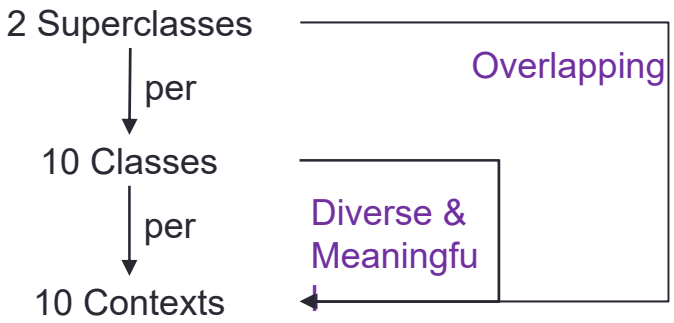


A dataset for Non-I.I.D. image classification is demanded.

NICO - Non-I.I.D. Image Dataset with Contexts

- **NICO** Datasets:
- Object label: e.g. dog
- Contextual labels (Contexts)
  - the background or scene of a object, e.g. grass/wate
- Structure of NICO





NICO - Non-I.I.D. Image Dataset with Contexts

Animal	Data Size	*Vehicle*	Data Size
Bear	1609	Airplane	930
Bird	1590	Bicycle	1639
Cat	1479	Boat	2156
Cow	1192	Bus	1009
Dog	1624	Car	1026
Elephant	1178	Helicopter	1351
Horse	1258	Motorcycle	1542
Monkey	1117	Train	750
Rat	846	Truck	1000
Sheep	918		

- Data size of each class in NICO
  - Sample size: thousands for each class
  - Each superclass: 10,000 images
  - Sufficient for some basic neural networks (CNN)
- Samples with contexts in NICO

Controlling NI on NICO Dataset

- Minimum Bias (comparing with ImageNet)
- Proportional Bias (controllable)
  - Number of samples in each context
- Compositional Bias (controllable)
  - Number of contexts that observed

Minimum Bias

- In this setting, the way of random sampling leads to minimum distribution shift between training and testing distributions in dataset, which simulates a nearly i.i.d. scenario.
  - 8000 samples for training and 2000 samples for testing in each superclass (ConvNet)

	Average NI	Testing Accuracy
Animal	3.85	49.6%
Vehicle	3.20	63.0%

Average NI on ImageNet: 2.7

Images in NICO are with rich contextual information

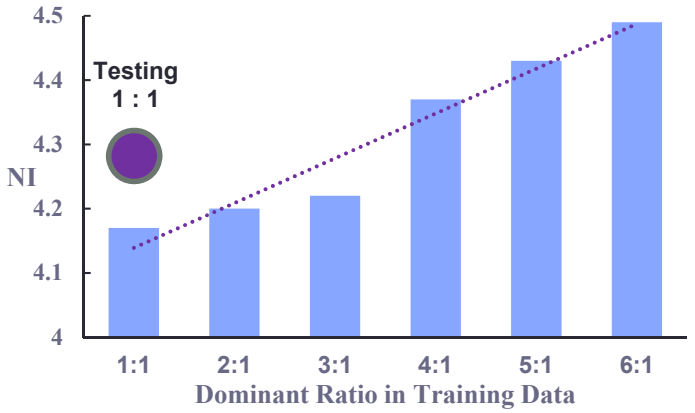
more challenging for image classification

Our NICO data is more Non-iid, more challenging

Proportional Bias

- Given a class, when sampling positive samples, we use all contexts for both training and testing, but the percentage of each context is different between training and testing dataset.

$$Dominant\ Ratio = \frac{N_{dominant}}{N_{minor}}$$

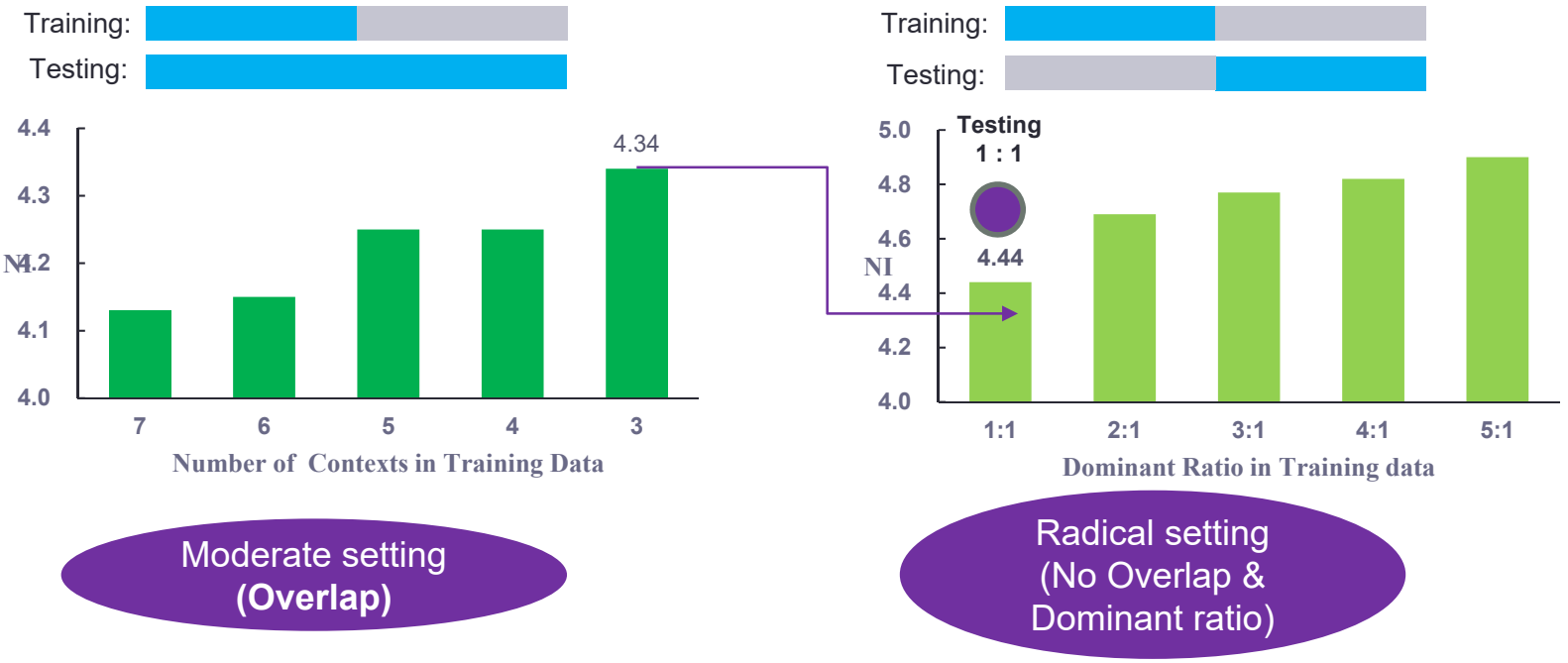


We can control NI by varying dominate ratio

Compositional Bias

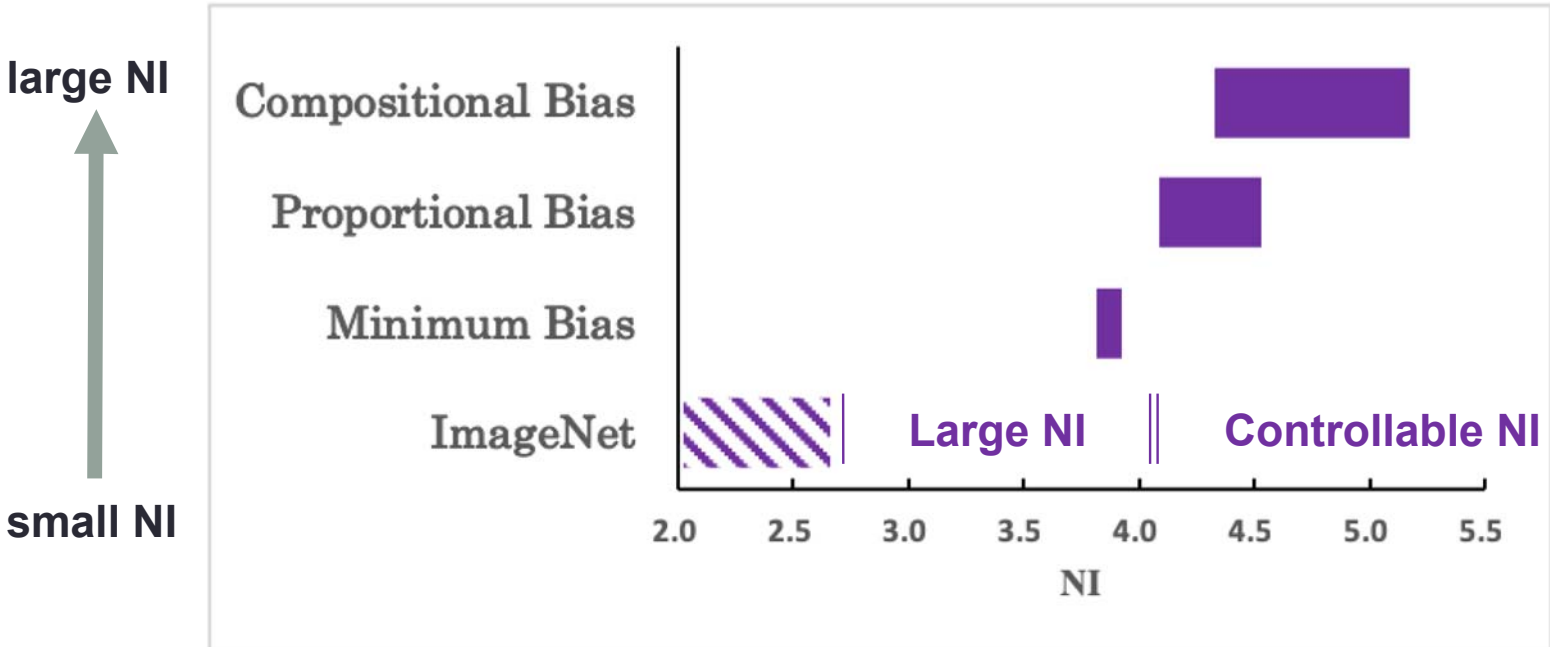
$$Dominant\ Ratio = \frac{N_{dominant}}{N_{minor}}$$

- Given a class, the observed contexts are different between training and testing data.



NICO - Non-I.I.D. Image Dataset with Contexts

- Large and controllable NI



NICO - Non-I.I.D. Image Dataset with Contexts

- The dataset can be downloaded from (temporary address):
- https://www.dropbox.com/sh/8mouawi5guaupyb/AAD4fdySrA6fn3PgSmhKwFgva?dl=0

- Please refer to the following paper for details:
- Yue He, Zheyan Shen, Peng Cui. NICO: A Dataset Towards Non-I.I.D. Image Classification. https://arxiv.org/pdf/1906.02899.pdf

Conclusions

- Beyond parameter tuning, it is more important to think about the learning mechanism.
- Human-like learning and reasoning is the valuable source to get inspirations.
- From black-box prediction models to explainable learning and reasoning processes is more meaningful.

Reference

- Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, Qi Tian. Learning to Learn Image Classifiers with Visual Analogy, ***CVPR*** 2019.
- Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.
- Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM Multimedia****, 2018*.
- Kun Kuang, Peng Cui, Bo Li, Shiqiang Yang. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. ***KDD***, 2017.
- Kun Kuang, Peng Cui, Bo Li, Shiqiang Yang. Treatment Effect Estimation with Data-Driven Variable Decomposition. ***AAAI***, 2017.
- Yue He, Zheyan Shen, Peng Cui. Towards Non-I.I.D. Image Classification: A Dataset and Baselines. (under review)
- Zheyan Shen, Peng Cui, Tong Zhang. Stable Learning of Linear Models via Sample Reweighting. (under review)

Thanks!

Peng Cui
cuip@tsinghua.edu.cn
http://media.cs.tsinghua.edu.cn/~multimedia/cuipeng/